The Biometric Society

BIOMETRICS

FOUNDED BY THE BIOMETRICS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 17: 1961

Numbers 1–4

Published Quarterly by
THE BIOMETRIC SOCIETY
OAK RIDGE, TENNESSEE
1961

BIOMETRICS

Published quarterly by the BIOMETRIC SOCIETY in March
June, September, and December

EDITORIAL AND BUSINESS OFFICE OF BIOMETRICS: Department of Statistics, Florida State University, Tallahassee, Florida, U. S. A. Manuscripts, non-member subscriptions, changes of address for non-member subscriptions, and undelivered copies should be sent to this office.

BUSINESS OFFICE OF THE SOCIETY: 101 Moore Lane, Oak Ridge, Tennessee, U.S.A. Members of the Biometric Society are advised to send changes of address and all correspondence regarding members' subscriptions to their Regional or National Secretary or, for members-at-large to the business office of the Society, P. O. Box 2017, Oak Ridge, Tennessee, U.S.A.

Second-class mailing privileges authorized at Richmond, Virginia.

OFFICERS OF THE BIOMETRIC SOCIETY

GENERAL OFFICERS

President	*Secretary*	*Treasurer*
Leopold Martin	M. J. R. Healy	M. A. Kastenbaum

COUNCIL

1959–1961	1960–1962	1961–1963
M. S. Bartlett, *BR*	G. S. Watson, *ENAR*	A. W. Kimball, *ENAR*
D. G. Chapman, *WNAR*	C. I. Bliss, *ENAR*	H. N. Turner, *AR*
C. W. Emmens, *AR*	D. J. Finney, *BR*	S. C. Pearce, *BR*
F. G. Fraga, *R. Bras.*	A. Linder, *Switzerland*	J. L. Hodges, *WNAR*
A. Lenger, *R. Belg.*	P. V. Sukhatme, *R. Ital.*	W. U. Behrens, *DR*
C. C. Li, *ENAR*	G. Teissier, *RF*	H. L. LeRoy, *Switzerland*
	F. Yates, *BR*	

REGIONAL OFFICERS

Region	*President*	*Secretary*	*Treasurer*
Australasian	M. Belz	W. B. Hall	G. W. Rogerson
Belgian	M. Welsch	L. Martin	P. Gilbert
Brazilian	A. M. Penha	E. Berquó	A. Groszmann
British	J. A. Fraser Roberts	C. D. Kemp	P. A. Young
E. N. American	Oscar Kempthorne	Erwin L. LeClerg	Donald A. Gardiner
French	Ph. L'Heritier	Sully Ledermann	Sully Ledermann
German	O. Heinisch	R. Wette	M. P. Geppert
Italian	G. Montalenti	R. Scossiroli	F. Sella
W. N. American	W. Taylor	W. Becker	Bernice Brown

NATIONAL SECRETARIES

Denmark	N. F. Gjeddebæk	*Netherlands*	H. de Jonge
India	A. R. Roy	*Norway*	L. K. Strand
Japan	M. Hatamura	*Sweden*	H. A. O. Wold
	Switzerland	H. L. LeRoy	

SUSTAINING MEMBERS OF THE BIOMETRIC SOCIETY

Abbott Laboratories
American Cancer Society, Inc.
General Foods Corporation, Research Laboratories
Merck, Sharp and Dohme Research Laboratories
Schering Corporation
Smith, Kline and French Laboratories
E. R. Squibb and Sons
The Upjohn Company
Wallace Laboratories, Division of Carter Products
Wyeth Institute of Applied Biochemistry

CONTENTS OF VOLUME 17

INDEX, BIOMETRICS VOLUME 17

The Biometric Society

BIOMETRICS

FOUNDED BY THE BIOMETRICS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION

TABLE OF CONTENTS

Articles

Queries and Notes

Book Reviews

Number 1 March 1961 Volume 17

BIOMETRICS

Officers of the Biometric Society

General Officers

President: Leopold Martin

Secretary: M. J. R. Healy; *Treasurer*: M. A. Kastenbaum

Council: M. S. Bartlett, W. U. Behrens, C. I. Bliss, A. Buzzati-Traverso, D. G. Chapman, W. J. Dixon, C. W. Emmens, D. J. Finney, C. G. Fraga, Jr., Anne Lenger, C. C. Li, A. Linder, K. Mather, C. R. Rao, P. V. Sukhatme, G. Teissier, J. W. Tukey, G. S. Watson, F. Yates.

Regional Officers

Region	*President*	*Secretary*	*Treasurer*
Australasian	M. Belz	W. B. Hall	G. W. Rogerson
Belgian and Congo	J. Henry	L. Martin	A. H. L. Rotti
Brazilian	A. M. Penha	A. Conagin	A. Groszmann
British	J. A. Fraser Roberts	C. D. Kemp	P. A. Young
E. N. American	Oscar Kempthorne (Henry L. Lucas)	Erwin L. LeClerg	Donald A. Gardiner
French	Ph. L'Heritier	D. Schwartz	D. Schwartz
German	O. Heinisch	R. Wette	M. P. Geppert
Italian	G. Montalenti	R. Scossiroli	F. Sella
W. N. American	W. A. Taylor	W. A. Becker	Bernice Brown

National Secretaries

Denmark	N. F. Gjeddebæk	*Netherlands*	H. de Jonge
India	A. R. Roy	*Norway*	L. K. Strand
Japan	M. Hatamura	*Sweden*	H. A. O. Wold
	Switzerland	H. L. LeRoy	

MARGINAL PERCENTAGES IN MULTIWAY TABLES OF QUANTAL DATA WITH DISPROPORTIONATE FREQUENCIES

F. YATES
Rothamsted Experimental Station,
Harpenden, Herts., England

SUMMARY

When multiway tables of quantal data are analysed by transforming the percentages and fitting constants to the transformed variate by the method of maximum likelihood, direct inverse transformation of the values of the fitted constants gives percentages which may deviate widely from those observed. The reasons for this are explained and methods are given for obtaining correct percentages.

INTRODUCTION

Transformations, such as the logit and log log, are now being used increasingly for the analysis of multiway tables of quantal data. The basic procedures have been described by Jolly [1950], Dyke and Patterson [1952] and Yates [1955]. Their earlier use for this purpose was undoubtedly inhibited by the heavy numerical computation required, but the burden of this has been greatly lightened by the introduction of electronic computers. The Rothamsted computer, for example, has a programme for analyses of this type which will handle tables with up to five factors in which the number of cells and marginal means does not exceed 256. Constants representing the main classifications and any desired sets of sub-classifications can be fitted, and the values of the constants and the reduction in variance are determined.

When an analysis of this kind has been completed, it is frequently desirable to present the results in terms of percentages, instead of, or as well as, in terms of the transformed variate. If the constants obtained in the analysis are retransformed directly to percentages, considerable distortions may occur. It is the purpose of this paper to explain why such distortions arise and how they may be removed.

ESTIMATION OF A MEAN PERCENTAGE BY MEANS OF A TRANSFORMATION

It is well known that, if a set of percentages is transformed by a non-linear transformation such as the logit transformation, and the mean of the transformed values is retransformed back to a percentage, this percentage will not equal the mean of the original percentages. The ordinary maximum likelihood procedure of fitting, using provisional and working values and successive approximation, removes this distortion.

This result follows immediately from the fact that the weighted mean of a set of percentages, with weights proportional to the frequencies, is the maximum likelihood estimate of the mean percentage. A direct proof may, however, be of interest, and illustrates the procedure to be followed in the more complicated case of two-way tables, discussed in the next section.

Suppose we have a set of proportions $p_r = x_r/n_r$, where n_r is the number of observations in class r and x_r the number of those fulfilling the required condition, so that 100 p_r is the corresponding percentage.

If the observations in all the classes are in fact samples from the same population, the sufficient estimate of the population proportion is

$$\hat{p} = \sum x_r / \sum n_r = \sum n_r p_r / \sum n_r .$$

If, with any transformation $y = f(p)$, with inverse $p = F(y)$, a common provisional value Y is taken, and Y_{max} and Y_{min} are the maximum and minimum values and w the weighting coefficient for this value of Y (as defined in the ordinary maximum likelihood procedure), then the working value for class r will be

$$y_r = p_r Y_{\text{max}} + q_r Y_{\text{min}} .$$

The weighted mean of these values, which will provide a first estimate of the adjusted mean, and also a new provisional value, will be

$$\begin{aligned} Y' &= \frac{\sum n_r w y_r}{\sum n_r w} = \frac{\sum n_r p_r}{\sum n_r} Y_{\text{max}} + \frac{\sum n_r q_r}{\sum n_r} Y_{\text{min}} \\ &= \hat{p} Y_{\text{max}} + \hat{q} Y_{\text{min}} . \end{aligned}$$

If successive approximations are now carried out until there is no change in the provisional value, we shall arrive finally at the equations

$$\begin{aligned} Y &= \hat{p} Y_{\text{max}} + \hat{q} Y_{\text{min}} , \\ \hat{y} &= Y, \end{aligned} \tag{1}$$

where Y, Y_{max} and Y_{min} are the final values of Y' and the corresponding maximum and minimum. If $P = F(Y)$, the formulae for Y_{max} and Y_{min} are

$$Y_{\max} = Y + Q\frac{dY}{dP},$$

$$Y_{\min} = Y - P\frac{dY}{dP}.$$

(See, for example, Fisher and Yates [1957], p. 14.) Hence equation (1) reduces to $\hat{p}Q - \hat{q}P = 0$, i.e. $P = \hat{p}$, and therefore $\hat{p} = F(\hat{y})$. In other words the inverse transformation of the maximum likelihood estimate $\hat{y}$ gives an estimate $\hat{p}$ which is equal to the weighted mean of the original percentages, with weights proportional to the frequencies.

TWO-WAY TABLES

In the analysis of a multi-way table of quantal data the expected values of the transformed variate are assumed to be made up of additive components representing the various classifications which require to be taken into account.

Thus in an $R \times S$ two-way table the expected value of the transformed variate for the cell (r, s) is given by

$$\hat{y}_{rs} = m + a_r + b_s,$$

where m is a general mean, $a_1, \cdots, a_r$ is the set of constants for the first classification, and $b_1, \cdots, b_s$ the set for the second classification. The values of m and the a's and b's are obtained by the maximum likelihood fitting.

If the $\hat{y}_{rs}$ are transformed back to percentages $100\,\hat{p}_{rs}$, then $100\,\hat{p}_{rs}$ will represent the expected percentages on the assumed hypothesis without distortion. Examination of the equations which are used to determine the fitted constants, on the lines of the last section, shows that in the case of the logit transformation the weighted means of $\hat{p}_{rs}$, with weights equal to the original frequencies, will reproduce the marginal percentages of the original table exactly. This depends on the fact that for this transformation

$$w\frac{dY}{dP} = \text{const},$$

which follows from the general formula for w, namely

$$w = \frac{1}{PQ}\left(\frac{dP}{dY}\right)^2.$$

With other transformations there will be some discrepancy. There is, however, no real inconsistency; the marginal percentages so obtained, not those observed, are the appropriate estimates of these percentages for the adopted model, with the observed weights.

To represent the estimates of the class-differences in terms of percentages in compact form, we require the expected percentages of the margins of the table with some form of proportionate weights. As in the quantitative case it is usually best to give the percentages that would be expected if the frequencies in the cells of the table were proportionate to the marginal frequencies, i.e.

$$n'_{rs} = n_{r.}n_{.s}/n_{..} \; ,$$

where

$$n_{r.} = \sum_{s} n_{rs} \; , \quad \text{etc.}$$

These percentages can be computed directly from the percentages 100 $\hat{p}_{rs}$ with the appropriate weights, e.g.

$$\begin{aligned}\hat{p}_{r.} &= \sum_{s} \hat{p}_{rs}n'_{rs}/n_{r.} \; , \\ &= \sum_{s} \hat{p}_{rs}n_{.s}/n_{..} \; .\end{aligned}$$

We will refer to this procedure as computation from the expected cell percentages.

Computation of the expected marginal percentages from the expected cell percentages, if done by hand, is somewhat tedious, particularly in large multiway tables, since the expected percentages of the individual cells have first to be computed and their weighted means then taken. The question therefore arises whether the values of the fitted constants, which are given directly by the computation, can be used to provide values of expected marginal percentages.

In the analogous analysis of quantitative data the values of the fitted constants give direct estimates of the values the marginal means would assume if the cell frequencies were proportionate to the marginal frequencies. With exactly proportionate cell frequencies the values of fitted constants $m + a_r$ and $m + b_s$ are given directly by the two sets of marginal means. In the case of quantal data, however, retransformation of the values of the fitted constants $m + a_r$ and $m + b_s$ will give percentages which may differ widely from those obtained by computation from the expected cell percentages.

These differences are due to two causes. In the first place, since the fitted constants are in effect weighted means of the expected cell values of the transformed variates, the corresponding percentages will differ from the means, with the same weights, of the cell percentages.

In the second place the weights used in the fitting are the products of the cell frequencies and the weighting coefficients. The resultant fitted constants are therefore estimates of what the marginal means of the transformed variate would be if the weights were proportionate to the marginal sums of these weights instead of to the marginal frequencies. If the angular transformation, which has a constant weighting coefficient, is used, no differences will arise from this cause, but with other transformations it is frequently the major source of discrepancy.

At first sight it might appear that the best course would be to seek a value m' such that the quantities $m' + a_r$ when transformed give percentages whose weighted mean, using the marginal frequencies as weights, is equal to the observed overall percentage. This can be done without difficulty by successive approximation, and appears to be reasonably satisfactory with the angular transformation. When the percentages are small and the logit transformation is used, a much closer approximation to the marginal percentages computed from the expected cell percentages is obtained by the very simple procedure of transforming $m + a_r$ to percentages and multiplying all these percentages by a common factor, chosen so as to give the observed overall percentage, as is illustrated in the example below. This is to be expected since at the lower end of the percentage scale the use of the logit transformation implies that the cell percentages are representable by a product function of the form $u_r v_s$.

It may be noted here that even with the logit transformation the overall percentage computed from the expected cell percentages will differ from the observed overall percentage, owing to the change from actual to proportionate frequencies in the cells. The difference is not likely to be large, but for practical purposes it may be regarded as preferable to present marginal percentages whose weighted means are exactly equal to the observed overall percentage. For this purpose a proportional adjustment of all percentages (or their differences from 100 when all percentages are large) will be adequate.

EXAMPLE

Table 1 gives the percentages of cows affected by milk fever in given lactations, classified by lactation number and season of calving, and Table 2 the numbers of calvings on which these percentages are based. These results were obtained in the course of a survey of diseases of dairy cattle.

It is immediately apparent from the percentages that there is a much greater incidence of milk fever in the later calvings, and also a variation with season, January-April having the lowest incidence and

TABLE 1

Percentages of cows affected by milk fever

	Lactation					
	1–2	3	4	5	6+	Overall
Jan–Apr	0.42	1.45	3.15	5.57	6.07	2.67
May–July	0.17	3.08	7.37	9.43	10.94	4.17
Aug–Sept	0.66	4.89	9.40	10.93	13.98	4.25
Oct–Dec	0.35	2.93	6.07	9.61	12.19	3.50
Overall	0.41	2.75	5.58	8.06	9.38	3.46

August-September the highest. The August-September maximum is somewhat masked in the marginal percentages, however, owing to the greater proportions of these calvings in the earlier lactations.

Constants were fitted using the logit transformation with logs to base 2. The results shown in Table 3 were obtained. Inverse transformation of these logits gives percentages for seasons which are very much too high. This is mainly due to the low weighting coefficients associated with the low percentages of lactations $1 - 2$.

The expected logits for the individual cells can be computed from the values of Table 3, and transformed to percentages. Thus the logit for the top left-hand cell is $-4.0121 - 2.3786 + 1.9874 = -4.4033$, and the corresponding expected percentage is 0.2229. The discrepancies between the weighted marginal means of these percentages, with weights equal to the original cell frequencies (Table 2), and the observed marginal percentages, are all less than 0.001.

TABLE 2

Numbers of Calvings

	Lactation					
	1–2	3	4	5	6+	Total
Jan–Apr	3806	2137	1744	1184	2010	10881
May–July	2352	1006	706	488	841	5393
Aug–Sept	3169	1003	617	366	522	5677
Oct–Dec	5117	1740	1252	791	1042	9942
Total	14444	5886	4319	2829	4415	31893

TABLE 3

Fitted Constants (Logits to Base 2) and Corresponding Percentages

Lactation	$m + a_r$	%	Season	$m + b_s$	%
1–2	−4.0121	0.383	Jan–	−2.3786	3.566
3	−2.5643	2.779	May–	−1.9180	6.544
4	−2.0055	5.840	Aug–	−1.6682	9.009
5	−1.7082	8.564	Oct–	−1.9104	6.609
6+	−1.5678	10.220	—	—	—
m	−1.9874	5.980			

Table 4 gives the various estimates of the expected marginal percentages with proportionate weights. Line (1) gives the weighted means of the expected cell percentages, with weights equal to the marginal frequencies of Table 2. These give a small discrepancy in the overall percentage, and this is removed in line (2) by proportional adjustments. Line (3) is obtained by proportional adjustment of the percentages derived directly from the fitted constants (Table 3). The agreement

TABLE 4

Marginal Percentages Computed by Various Methods

(1) From expected cell percentages, with proportionate weights
(2) Line (1) adjusted to observed overall percentage
(3) From fitted constants (adjusted)

Lactation

	1-2	3	4	5	6+	Overall
(1)	0.39	2.79	5.85	8.55	10.18	3.65
(2)	0.37	2.65	5.55	8.11	9.66	3.46
(3)	0.36	2.63	5.54	8.12	9.69	3.46

Season

	Jan	May	Aug	Oct	Overall
(1)	2.19	3.99	5.46	4.03	3.65
(2)	2.08	3.78	5.18	3.82	3.46
(3)	2.06	3.78	5.21	3.82	3.46

between lines (2) and (3) is very close. With the logit transformation and small percentages direct derivation from the fitted constants may therefore be regarded as quite satisfactory.

Comparison of these expected percentages with the observed percentages of Table 1 shows that the distortions in the observed marginal percentages due to disproportionate weights have been removed by the fitting.

The residual sum of squares given by the fitting is 17.57. This corresponds to a χ^2 with 12 *d.f.*, giving $.2 < P < .1$. The data are therefore reasonably represented by additive components in the logit scale.

MULTI-WAY TABLES

If only constants representing the main classifications are fitted the procedure of proportional marginal adjustment is exactly the same for three- or more-way tables as for two-way tables. If, however, constants representing certain sets of sub-classes (analogous to the interactions of a factorial experiment) are also fitted, the situation becomes more complex. Nevertheless if the expected values of the transformed variate are available only for the margins of the table, a procedure analogous to that given above for a two-way table will probably be quite satisfactory, but certain minor inconsistencies in the margins relating to the main classifications may be expected.

PROGRAMMING POINTS IN ELECTRONIC COMPUTATION

Any programme for the analysis of multiway tables of quantal data which follows the ordinary maximum likelihood procedure of successive approximation must embody provisions for computing the expected values, in terms of the transformed variate, of the individual cells of the table, since these values constitute the provisional values for the next approximation. It must also contain provision for forming the weighted marginal means, with disproportionate and probably also with proportionate weights. Consequently, in order to calculate the expected cell percentages all that is necessary, in the way of additions to the programme, is to re-calculate the expected values of the individual cells after the last approximation (this will probably be done in any case), transform these to percentages, and calculate the marginal means of these percentages with weights proportional to the marginal frequencies of the main classifications. If desired a proportional adjustment can be made to these percentages so as to give the observed overall percentage.

This procedure is likely to be just as simple, from the programming point of view, as the alternative procedure of adjusting the marginal percentages proportionally and should therefore be adopted when writing a complete programme. The latter procedure is, however, of interest, since it makes it possible to deal expeditiously with results furnished by a programme for which no provision for inverse transformation has been made, or which only inversely transforms the expected marginal values of the transformed variate.

REFERENCES

Dyke, G. V. and Patterson, H. D. [1952]. Analysis of factorial arrangements when the data are proportions. *Biometrics 8*, 1–12.

Fisher, R. A. and Yates, F. [1957]. *Statistical tables for biological, agricultural and medical research.* 5th Edition. Edinburgh: Oliver & Boyd.

Jolly, G. M. [1950]. Use of probits in combining percentage kills. *Ann. Appl. Biol. 37*, 597–606.

Yates, F. [1955]. The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments. *Biometrika 42*, 382–403.

SOME CLASSIFICATION PROBLEMS WITH MULTIVARIATE QUALITATIVE DATA[1]

WILLIAM G. COCHRAN
*Department of Statistics, Harvard University,
Cambridge, Massachusetts, U.S.A.*

AND

CARL E. HOPKINS
*Department of Biostatistics, University of Oregon Medical School,
Portland, Oregon, U.S.A.*

1. INTRODUCTION

Since 1935, when Fisher's discriminant function appeared in the literature, methods for classifying specimens into one of a set of universes, given a series of measurements made on each specimen, have been extensively developed for the case in which the measurements are continuous variates. This paper considers some aspects of the classification problem when the data are qualitative, each measurement taking only a finite (and usually small) number of distinct values, which we shall call *states*. Our interest in the problem arose from discussions about the possible use of discriminant analysis in medical diagnosis. Some diagnostic measurements, particularly those from laboratory tests, give results of the form: −, + (2 states); or −, doubtful, + (3 states); or (with a liquid), clear, milky, brownish, dark (4 states).

With qualitative data of this type an optimum rule for classification can be obtained as a particular case of the general rule (Rao, [1952], Anderson, [1958]). The rule is exceedingly simple to apply (Section 2). In practice, qualititative data are frequently ordered, as with −, doubtful, +. The classification rule discussed in this paper takes no explicit advantage of the ordering, as might be done, for instance, by assigning scores to the different states so as to produce quasi-continuous data. The best method of handling ordered qualitative data is a subject worth future investigation.

[1] This work was assisted by a contract between the Office of Naval Research, Navy Department, and the Department of Statistics, Harvard University by a Special Research Fellowship of the National Institutes of Health, and by a gift from the Lilly Research Laboratories to the Department of Biostatistics, University of Oregon Medical School. Reproduction in whole or in part is permitted for any purpose of the United States Government.

This paper seeks answers to three problems that arise in the use of the proposed rule.

(1) *The effect of the initial sample sizes on the performance of the proposed classification rule.*

In order to construct a rule, we must have preliminary data on some specimens known to be classified correctly, since the rule depends on the joint frequency distributions of the measurements within each universe. Standard classification theory assumes that these distributions are known exactly, although in practice it is sometimes difficult to obtain adequate samples from which to estimate them. The consequences of constructing a rule from preliminary samples of finite sizes are discussed in Sections 3, 4, and 5.

(2) *The relative discriminating power of qualitative and continuous variates.*

Classification is simpler with qualitative than with continuous measurements. For this reason, if a few of the available measurements are continuous while the rest are qualitative, we may be inclined to transform each continuous measurement into a qualitative one by partitioning its frequency distribution, provided that this does not result in too much loss of discriminating power. This consideration led us to investigate the questions: if a normally distributed variate is transformed into a qualitative one with s states, how much discriminating power is lost, and what are the best points of partition of the curve from the point of view of retaining maximum discriminating power (Sections 6 and 7)?

(3) *Use of classification experience for improvement of the rule.*

The optimum rule depends on the relative frequencies π_1 , π_2 , π_3 etc. with which specimens from the different universes present themselves for classification. The initial estimates of these frequencies, made from previous data or by judgement, may be biased. After a number of specimens have been classified by the rule, it is possible to re-estimate the frequencies π_u from the data for these specimens, so that the classification rule may be improved (Section 8).

2. AN OPTIMUM RULE FOR CLASSIFICATION

Ideally, the setting up of an optimum rule demands three different kinds of preliminary information. Firstly, we require the joint frequency distribution of the measurements in each universe under consideration. With continuous variates the most common assumption about these distributions is that they are multivariate normal with the same dispersion matrix in each universe. With qualitative variates, suppose that the jth measurement takes s_j distinct states. When this measurement is made on a specimen, we learn into which state it falls. A series of qualitative measurements, e.g. one with 2 states, one with 3 states and one with 4 states, defines 24 cells or multivariate states. For any specimen, these three measurements tell us which of the 24 states the specimen occupies. Thus, in general, a set of k qualitative measurements classifies the specimen into one out of a number $S = s_1 s_2 \cdots s_k$ multivariate states. Consequently the joint frequency distribution of the measurements is completely specified for the uth universe if we know the probabilities p_{ui} that the specimen falls into the ith multivariate state where, for each u,

$$\sum_{i=1}^{S} p_{ui} = 1.$$

Secondly, practical use of a classification rule is likely to result in some mistakes in classification. In some applications certain types of mistakes are more serious than others. If the relative costs of different kinds of mistakes can be estimated, the rule should take them into account. Let c_{vu} be the cost incurred when a specimen which actually belongs to universe V is classified as belonging to universe U.

Thirdly, when the rule is put into use, the specimens to be classified may not come with equal frequency from the universes. For instance, in a diagnostic screening test, most of the subjects who present themselves may be free from the disease in question. Let π_u denote the relative frequency with which specimens come from the uth universe, where

$$\sum_{u} \pi_u = 1.$$

As might be expected, the optimum rule depends on these frequencies.

With this background we now construct an optimum rule. For a specimen which falls into the ith multivariate state, we select the universe to which it is to be assigned by minimizing the expected cost of mistakes in classification. Suppose that the specimen is assigned to the uth universe. For specimens that actually come from this universe, no mistake is made. Specimens from the vth universe will present themselves with relative frequency π_v , and will fall into the ith state with relative frequency $\pi_v p_{vi}$. The expected cost of misclassifying these specimens into the uth universe is therefore $\pi_v p_{vi} c_{vu}$. Hence the total expected cost of mistakes in classification is

$$\sum_{v \neq u} \pi_v p_{vi} c_{vu} .$$

To apply the optimum rule we compute this quantity for every u and assign specimens to the universe for which the quantity is a minimum.

There are several particular cases of the rule. If $c_{vu} = c_v$, i.e. the cost of misclassifying depends only on the universe from which the specimen comes and not on that into which it is misclassified, the total expected cost over all m universes is

$$\sum_{v \neq u} \pi_v p_{vi} c_v = \sum_{v=1}^{m} \pi_v p_{vi} c_v - \pi_u p_{ui} c_u .$$

Since the first term on the right does not depend on u, specimens falling into the ith state are assigned to the universe for which the triple product $\pi_u p_{ui} c_u$ is greatest.

If the relative costs c_u are taken as equal, the rule minimizes the expected frequency of mistakes in classification. If, further, the relative frequencies π_u are also equal, any specimen found in the ith state is assigned to the universe u for which the conditional probability p_{ui} is greatest.

For a numerical example we are indebted to Dr. Leslie Kish. The data come from a large study of voting behavior conducted by the Survey Research Center, University of Michigan. (Stokes *et. al.* [1958]). By open-end questionnaires taken during the 1952 and 1956 elections, voters were rated as Democrats (D), Independents (I) or Republicans (R) and also as Unfavorable (U), Neutral (N) or Strongly Favorable (F) to Eisenhower's personality. These are the two predictor measurements, each with 3 states, making 9 bivariate states. The voters were also asked whether they voted for Stevenson or Eisenhower, these being the two universes. The sample sizes were 1003 and 1439.

For illustrative purposes we have set up a rule classifying the subjects as Stevenson or Eisenhower voters by means of the predictor variates. The results of the classification are then compared with the actual voting behavior.

Table 1 shows the relative frequencies p_{1i} and p_{2i} of the ith state for Stevenson and Eisenhower voters respectively. We assumed that $c_1 = c_2$ and that $\pi_1 = \pi_2 = \frac{1}{2}$. The resulting classification rule is given on the right in Table 1. Persons falling into the states DU, DN, DF and IU are classified by the rule as Stevenson voters, since in these states $p_{1i} > p_{2i}$.

The estimated probabilities of misclassification are easily obtained. Any Eisenhower voter is misclassified if he falls into the states DU, DN, *DF* or *IU*. Hence the frequency of misclassification for Eisenhower voters is

$$.033 + .079 + .090 + .017 = .219.$$

The figure for Stevenson voters is 0.133 and the average is 0.176.

As with continuous measurements, it is possible to examine whether a measurement contributes to the classification to a worthwhile extent. If political affiliation alone (*D*, *I*, *R*) is recorded, the probabilities in the three states are given in the lower part of Table 1, being obtained of course by addition from the upper part of the table. The probabilities of misclassification are now 0.201 for Stevenson voters and 0.202 for Eisenhower voters. Thus the addition of the attitude measurement reduces the average number of mistakes by about 13 percent.

The use of qualitative variates in attempting to date certain of the

TABLE 1

Illustration of the Classification Rule for Two Measurements

State (i)	p_{1i} Stevenson	p_{2i} Eisenhower	Resulting classification rule
DU	.347	.033	S
DN	.359	.079	S
DF	.094	.090	S
IU	.068	.017	S
IN	.078	.107	E
IF	.026	.154	E
RU	.002	.017	E
RN	.023	.176	E
RF	.004	.327	E
	1.001	1.000	

Probabilities Using Political Affiliation Alone

State	Stevenson	Eisenhower	Classification
D	.800	.202	S
I	.172	.278	E
R	.029	.520	E

works of Plato (i.e. to arrange a number of universes in a time sequence) has been discussed by Cox and Brandwood [1959].

In the rest of this paper discussion will be confined unless otherwise mentioned to the case of two universes, with $c_1 = c_2 = 1$ and $\pi_1 = \pi_2 = \frac{1}{2}$.

3. EFFECTS OF THE FINITE SIZES OF THE INITIAL SAMPLES

The initial sample from each universe serves two purposes. It is used to set up the classification rule and to estimate the probabilities of misclassification for specimens from the two universes, so that we can decide whether the classification is accurate enough to be satisfactory.

Some notation will be needed. To avoid double subscripts, let U, U' denote the two universes, and p_i, p'_i the true probabilities that a specimen will fall into the ith multivariate state. Independent samples of sizes n, n' are drawn from the universes. The numbers of specimens falling into the ith state are r_i, r'_i, and the corresponding estimated probabilities are $\hat{p}_i = r_i/n$, $\hat{p}'_i = r'_i/n'$. Since the true p_i, p'_i are unknown, the actual classification rule places a specimen in U if $\hat{p}_i > \hat{p}'_i$, in U' if $\hat{p}_i < \hat{p}'_i$. If $\hat{p}_i = \hat{p}'_i$, the decision is made by tossing a fair coin.

For given sample sizes n, n' the values of $\hat{p}_i$ and $\hat{p}'_i$ will vary from sample to sample. Thus the actual classification rule and its performance may change from sample to sample. In describing the consequences of finite sample sizes, we usually present the average results over all initial samples of given sizes n, n'.

The principal consequences of the finite sample sizes are as follows.

1. The probability of misclassification as estimated from the samples is, of course, subject to a sampling error that in moderately large samples may be shown to be approximately of the binomial type.

2. On the average, taken over all pairs of samples from a given pair of universes, the actual probability of misclassification is *greater* than the theoretical optimum: i.e. than the probability that would hold if the true p_i, p'_i were used to construct the classification rule.

3. The average estimated probability of misclassification is *less* than the theoretical optimum probability. From 2, it is also less than the average actual probability.

In other words, finite sample sizes involve two types of penalty. Owing to sampling errors in the $\hat{p}_i$, $\hat{p}'_i$, the rule obtained may not be the theoretical optimum, and the probability of misclassification is underestimated from the sample data.

These results may be illustrated, in exaggerated form, by considering samples of size $n = 1$ from universes having two states with the following true probabilities of falling into the states. With the optimum rule the probability of misclassification is 0.1 for specimens from U, 0.05 for specimens from U', the average being 0.075.

State	U	U'	Optimum classification
1	0.9	0.05	U
2	0.1	0.95	U'

With $n = 1$, only four types of sample result are possible. Table 2 shows for each type the classification rule that would be set up, the estimated probability of misclassification, and the actual probability that would pertain if that rule were used. A ? means that classification is a toss-up.

To explain the entries in Table 2, consider the last type of result. The wrong classification is made in both states. Nevertheless, if we

TABLE 2

The Classification Rule for Samples of Size 1

Sample result				Frequency of occurrence	Probability of misclassification	
State	U	U'	Rule		Estimated	Actual
1	1	0	U	(0.9)(0.95)		
2	0	1	U'	=0.855	0.0	0.075
1	1	1	?	(0.9)(0.05)		
2	0	0	?	= 0.045	0.5	0.500
1	0	0	?	(0.1)(0.95)		
2	1	1	?	= 0.095	0.5	0.500
1	0	1	U'	(0.1)(0.05)		
2	1	0	U	= 0.005	0.0	0.925
				Average	0.07	0.13875

believe the samples, we estimate a zero probability of misclassification. The actual probabilities are 0.9 for U and 0.95 for U', giving an average of 0.925.

The overall average actual probability of misclassification is 0.13875 as compared with a theoretical optimum of 0.075 and an average estimated value of 0.07.

The fact that the average actual probability exceeds the theoretical optimum arises, of course, because when $p_i > p'_i$, there will be some pairs of samples in which $\hat{p}_i \leqq \hat{p}'_i$. Whenever this occurs, the actual rule for this state is not as good as the optimum.

The fact that the average estimated probability lies below the theoretical optimum probability may be shown as follows. If $p_i > p'_i$, the contribution of this state to the overall theoretical probability of misclassification is $\frac{1}{2}p'_i$. Since r'_i/n' is an unbiased estimate of p'_i, the contribution may be written $\frac{1}{2}E(r'_i/n')$, where E denotes a mean value. The estimated contribution, on the other hand, is $\frac{1}{2}$ the average of the *smaller* of r_i/n, r'_i/n', which will always be less than $\frac{1}{2}E(r'_i/n')$.

The bias in the estimated probability and the inferiority of the actual to the optimum rule are due primarily to states in which p_i differs little from p'_i. These difficulties can be largely avoided by withholding a decision, i.e. making no classification, in such states. For instance, suppose that preliminary samples of size 64 give the following results for a variate with four states.

	No. of specimens	
State	U	U'
1	51	1
2	8	5
3	3	5
4	2	53
	64	64

States 1 and 4 provide good discrimination, but in states 2 and 3 the estimated probabilities of a correct classification are only about 0.6. Consider a recommended rule to withhold judgement when a specimen occurs in state 2 or 3. From the samples, we have an unbiased estimate 21/128, or 16 percent, of the proportion of specimens for which the rule makes no classification. This estimate is binomial, with standard error $\sqrt{(0.16)(0.84)/128}$. When a decision is made, the probability of misclassification is estimated as 3/107, or about 3 percent. This estimate will be almost exactly a binomial variate, with negligible bias.

If, on the contrary, a decision is made in all states, we would like to obtain from the initial samples a reasonably unbiased estimate of the probability of misclassification for the actual rule developed from the samples. An estimate of the difference between the actual and the theoretical optimum probability of misclassification is also of interest, particularly when it is not too costly to increase the size of the initial samples. These topics are discussed in Sections 4 and 5.

4. ESTIMATION OF THE THEORETICAL OPTIMUM PROBABILITY

As shown in Section 3, the sample probability of misclassification is a negatively biased estimate of the theoretical optimum probability. For a given value of $(p_i + p'_i)$, it appears intuitively that the underestimation will be greatest when $p_i = p'_i$. In this case the bias in the ith state is in absolute value,

$$\tfrac{1}{2}p_i - \{\tfrac{1}{2}E \text{ min.} (\hat{p}_i , \hat{p}'_i)\}$$

where $\hat{p}_i$, $\hat{p}'_i$ are independent binomial estimates of p_i. An algebraic expression for this quantity as a function of p_i, can be developed from the binomial distribution, but is unwieldy unless n, n' are small. An approximation to the bias may be obtained by regarding $\hat{p}_i$, $\hat{p}'_i$ as normally distributed. Since the average value of the smaller of two normal deviates is known to be -0.56, the bias in the estimate of p_i when both samples are of size n is approximately

$$-0.56 \sqrt{p_i q_i / n}.$$

With $p_i = p'_i = 0.05$, $n = 50$, for instance, this estimate gives -0.0173 as against the correct value of -0.0170.

Since this problem will occur mainly in states with low frequencies (unless the classification rule is poor), a relatively crude sample adjustment for bias should be adequate for estimating the overall theoretical optimum probability of misclassification. One suggestion is to use $(\hat{p}_i + \hat{p}'_i)/4$ instead of one-half the smaller of $\hat{p}_i$, $\hat{p}'_i$ when estimating the probability of misclassification, this adjustment being made in states in which $\hat{p}_i$, $\hat{p}'_i$ do not differ significantly at, say, the 20 percent level of significance. In the numerical example above, the adjustment is made in states 2 and 3. It amounts to estimating the theoretical optimum probability as

$$\frac{1 + 6.5 + 4 + 2}{128} = \frac{13.5}{128} = 10.5\%$$

as compared with the unadjusted estimate of 11/128 or 8.6 percent.

Examination of individual cases indicates that this adjustment removes most of the bias in states in which $p_i = p'_i$. When $p_i \neq p'_i$ the adjustment tends to over-correct, and on the whole is likely to be conservative.

5. THE AVERAGE INCREASE IN THE ACTUAL PROBABILITY OF MISCLASSIFICATION

The average increase in the actual over the theoretical probability of misclassification may be expressed as a function of the p_i, p'_i and the sample sizes n (assumed equal). In the ith state, if $p_i > p'_i$, the optimum rule gives no misclassification for specimens from U, but misclassifies the proportion p'_i of specimens from U' that are in this state. The average frequency of misclassification is therefore $\frac{1}{2}p'_i$. With the samples, there is no increase in this frequency provided that $r_i > r'_i$. If $r_i = r'_i$, we toss a coin. This gives an increase in frequency of

$$\tfrac{1}{4}(p_i + p'_i) - \tfrac{1}{2}p'_i = \tfrac{1}{4}(p_i - p'_i).$$

If $r_i < r'_i$, we make the wrong decision, with an increase in frequency of mistakes of

$$\tfrac{1}{2}p_i - \tfrac{1}{2}p'_i = \tfrac{1}{2}(p_i - p'_i).$$

Hence, over all states for which $p_i > p'_i$, the total expected increase in frequency of misclassification may be written

$$\tfrac{1}{2} \sum (p_i - p'_i)\{\text{Pr.}\ (r_i < r'_i) + \tfrac{1}{2}\ \text{Pr.}\ (r_i = r'_i)\}, \qquad (5.1)$$

with an analogous expression for the states in which $p_i < p'_i$.

The probability inside the curly brackets may be obtained from tables of the binomial distribution. The normal approximation to this quantity is surprisingly good, even for moderate samples and small p_i, p'_i. With this approximation, the estimate is the probability that a normal deviate is less than

$$-\frac{\sqrt{n}\,|\,p_i - p'_i\,|}{\sqrt{p_i q_i + p'_i q'_i}} \qquad (5.2)$$

No correction for continuity is applied, because of the presence of the term $\frac{1}{2}Pr.\ (r_i = r'_i)$ in the exact expression. As an example of the accuracy, with $p_i = 0.03$, $p'_i = 0.01$, $n = 100$, the normal approximation gives 0.1555 as compared with the exact value 0.1567.

Use of the normal approximation is illustrated in Table 3. From the $\hat{p}_i$, $\hat{p}'_i$ for each state the normal deviate is computed from (5.2) and the probability is read from the normal tables. The difference

TABLE 3
ESTIMATION OF ACTUAL MINUS OPTIMUM PROBABILITY OF MISCLASSIFICATION

State	No. of specimens U	U'	N.D.	Probability	$\lvert \hat{p}_i - \hat{p}'_i \rvert$
1	51	1	Large	0.0	—
2	8	5	−0.880	0.189	0.0469
3	3	5	−0.732	0.232	0.0312
4	2	53	Large	0.0	—
Total	64	64			

between the average actual and the theoretical probability of misclassification is estimated as

$$\tfrac{1}{2}\{(0.189)(0.0469) + (0.232)(0.0312)\} = 0.0080 = 0.8\%.$$

At the end of Section 4 the theoretical probability of misclassification was estimated as 10.5 percent. Adding 0.8 percent, we estimate the expected actual probability as 11.3 percent.

In order to obtain a general impression of the effect of sample size, some study was made of expression (5.1) for the average increase of the actual over the theoretical optimum probability, as a function of p_i , p'_i , n, and S, the number of states. The difficulty in studying this function lies in the substantial number of parameters involved when the states are numerous. A preliminary search was made for a smaller number of parameters which might dominate the performance of the function. It became evident that the function depends on the overall optimum probabilities of misclassification, i.e. on the quantities

$$P = \sum_{p_i < p'_i} p_i \,, \qquad P' = \sum_{p'_i < p_i} p'_i \,.$$

As P, P' diminish, i.e. as the optimum rule becomes better, the average increase also diminishes, i.e. the actual rule is less likely to differ from the theoretical optimum. Even with the quantities P, P', n and S fixed, the average increase may vary from almost zero up to a maximum, depending on the way in which p_i , p'_i distribute themselves among the states. We decided to tabulate the maximum increase for given P, P', n and S, although aware that this might produce an unduly pessimistic view of the effect of finite sample size. The worst possible distributions of the p_i , p'_i were found by trial and error. In most situations, though not in all, the worst case occurs when the probability

P' is distributed equally among all but one of the states for which $p_i > p'_i$. As an example, for 8 states, with $P = P' = 0.1$, the worst distribution in the first four states is as follows.

p_i	p'_i
.69	.0
.07	.0333
.07	.0333
.07	.0333

Table 4 shows the maximum average increases for $n = 50, 100$; $S = 4, 8, 16, 32$; $P = P' = 0.05.\ 0.1, 0.2, 0.3, 0.4$.

TABLE 4

Maximum Average Increase in Probability of Misclassification

Opt. Prob.	$n = 50$				$n = 100$			
	Number of states				Number of states			
$P = P'$	4	8	16	32	4	8	16	32
0.05	.009	.018	.032	.059	.006	.011	.020	.034
0.1	.011	.022	.039	.067	.008	.015	.025	.041
0.2	.014	.027	.049	.082	.010	.020	.033	.052
0.3	.016	.034	.057	.098	.011	.023	.038	.062
0.4	.025	.037	.061	.071	.012	.026	.041	.061

As mentioned previously, the average increase is smallest when P, P' are small. It becomes larger as the number of states increases, being somewhat less than doubled for each doubling of the number of states. The average increases for $n = 100$ are roughly $1/\sqrt{2}$ times those for $n = 50$. Some irregularities in the table will be noticed for $P = 0.4$. The explanation is that the actual rule always gives at least 50 percent correct classifications. This imposes an upper limit to the maximum average increase for $P = 0.4$.

It is difficult to give general recommendations from this analysis. The following comments may be appropriate.

1. Preliminary samples of around 50 specimens from each universe should give a good indication of those multivariate states in which the classification will be accurate, and of those in which it will be dubious.

2. If a classification is to be made for all multivariate states, the estimated probabilities of misclassification should be increased by the two adjustments for bias.

3. If the number of multivariate states does not exceed 8, Table 4 suggests that samples of size 50 should make the actual rule satisfactorily close to the optimum in most situations. With 8 states and $n = 50$, the actual probabilities in the worst cases are 6.8, 12.2, 22.7 and 33.4 percent, for theoretical optimum probabilities of 5, 10, 20 and 30 percent, respectively. For larger numbers of states, substantially greater samples are indicated.

6. THE EFFECT OF REPLACING A NORMAL VARIATE BY A QUALITATIVE VARIATE

When some variates are continuous and some qualitative, Rao's general rule for classification still applies. However, it presents difficulties in application, as Linhart [1959] has pointed out, since a different continuous discriminant is required for each multivariate state defined by the qualitative variates.

In this section the loss of discriminating power when normally distributed variates are partitioned into qualitative variates is examined. It is assumed that the variates are independent. This assumption is unrealistic for applications, but we have not been able to reach general results with correlated variates. With independent variates, simple asymptotic results can be obtained when the number of variates becomes large, and can be checked against the exact results for a small number of variates.

Let the vth continuous variate be denoted by y_v, $(v = 1, \cdots, k)$, and the two universes by A and B. The variate y_v is normal, with s.d. unity in each universe, and with mean 0 in A and δ_v in B, the quantity δ_v being "the distance apart" of the two populations. The larger δ_v is, the better y_v is for classification.

Given k independent variates, it is easy to show that the best continuous discriminator, when divided so that its standard deviation is 1, is the quantity

$$R_c = \sum \delta_v y_v / \sqrt{\sum \delta_v^2} . \tag{6.1}$$

For this discriminator the distance between the two populations is $\sqrt{\sum \delta_v^2}$. As k becomes large with all $\delta_v > 0$, the distance becomes so great that classification is almost perfect. In order to discuss the situation in which some mistakes in classification are made, we assume that as k becomes large the δ_v tend to zero in such a way that $\sqrt{\sum \delta_v^2}$ remains finite.

The continuous classification rule assigns a specimen to B if $R_c \geq \frac{1}{2}\sqrt{\sum \delta_v^2}$. Since R_c is normally distributed, the probability of misclassification, which is the same for both universes, is the probability that a normal deviate is greater than $\frac{1}{2}\sqrt{\sum \delta_v^2}$.

Let each continuous variate be partitioned into a qualitative variate with six states. Results for fewer than six states are derived as particular cases of the result for six states. For the vth variate, let Pr. $(j_v \mid A)$, Pr. $(j_v \mid B)$ be the probabilities that a specimen from A and B respectively falls into the j_vth state. The 6^k multivariate states into which the specimen can fall may be specified by recording the states j_1, j_2, $\cdots$, j_k for the individual variates. Since the variates are independent, the probabilities that a specimen falls into a given multivariate state are

$$\prod_{v=1}^{k} \text{Pr. } (j_v \mid A) \quad \text{and} \quad \prod_{v=1}^{k} \text{Pr. } (j_v \mid B).$$

Consequently the optimum discrete rule places a specimen in A if

$$\prod_{v=1}^{k} \text{Pr. } (j_v \mid A) > \prod_{v=1}^{k} \text{Pr. } (j_v \mid B),$$

i.e. if

$$\sum_{v=1}^{k} \{\log \text{Pr. } (j_v \mid A) - \log \text{Pr. } (j_v \mid B)\} > 0. \tag{6.2}$$

This may be written

$$\sum x_v > 0 \tag{6.2$'$}$$

where

$$x_v = \log \text{Pr. } (j_v \mid A) - \log \text{Pr. } (j_v \mid B).$$

The quantity $\sum x_v$ is a discrete random variable with 6^k possible values, since its value depends on the multivariate state into which the specimen falls. It will be shown that as k becomes large, this quantity tends to be normally distributed.

The way in which an individual variate is partitioned is shown in Figure 1. The six states are labeled, for mnemonic reasons, A, a, α, β, b, B: they might be called "strongly, moderately, slightly favorable to A", etc. The partition involves two disposable numbers u_1, u_2, which determine the sizes of the regions α, β and a, b. (Four numbers might be used, but symmetry suggests that two will give the best partition.) Strictly, the numbers should be denoted by u_{1v}, u_{2v}, since it may be profitable to vary them with v, but for simplicity the subscript v will be omitted from u_1, u_2 and δ at present.

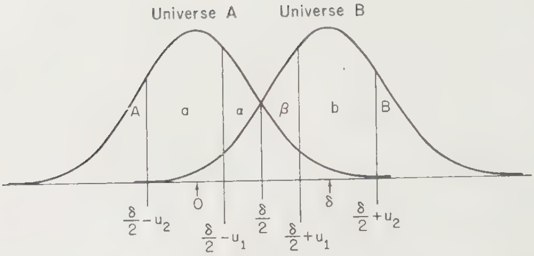


FIGURE 1

Partition of a Normal Variate Into Six States

For an individual variate v Table 5 shows the notation used for the probabilities of falling into the respective states.

The probabilities p_i are integrals over the parts of the normal curve indicated in Figure 1. Note that the Pr. ($|\ B$) values are the same as the Pr. ($|\ A$) values in reverse order.

TABLE 5

Probabilities That a Variate Will Fall Into the 6 States

	State					
	A	a	α	β	b	B
Pr. ($\|\ A$)	p_1	p_2	p_3	p_4	p_5	p_6
Pr. ($\|\ B$)	p_5	p_6	p_4	p_3	p_2	p_1
x	w_1	w_2	w_3	$-w_3$	$-w_2$	$-w_1$

The last line of Table 5 shows the six possible values taken by the random variable

$$x = \log \text{Pr.}\ (|\ A) - \log \text{Pr.}\ (|\ B)$$

where

$$w_1 = \log (p_1/p_6), \qquad w_2 = \log (p_2/p_5), \qquad w_3 = \log (p_3/p_4).$$

It follows from Table 5 that

$$E(x \mid A) = w_1(p_1 - p_6) + w_2(p_2 - p_5) + w_3(p_3 - p_4)$$

while $E(x \mid B)$ has the same value with sign reversed. The variance has the same value for both universes, namely

$$V(x) = w_1^2(p_1 + p_6) + w_2^2(p_2 + p_5) + w_3^2(p_3 + p_4) - [E(x \mid A)]^2.$$

Reverting to inequality (6.2)′, its left side is the sum of a large number of independent variables x, and hence tends to become normally distributed as k becomes large.

It remains to obtain the values of $E(x \mid A)$ and $V(x)$ when δ tends to zero. The leading terms in the expressions for the p_i and the w_i are found to be as follows.

Let

$$I_1 = \int_{u_1}^{\infty} z(t)\,dt, \qquad I_2 = \int_{u_2}^{\infty} z(t)\,dt$$

where $z(t)$ is the normal curve, and let z_0, z_1, z_2 be the ordinates of this curve at 0, u_1, u_2, respectively. Then, from Figure 1,

$$p_1 = \int_{-\infty}^{\frac{1}{2}\delta - u_2} z(t)\,dt \doteq I_2 + \tfrac{1}{2}\delta z_2\,,$$

$$p_2 = \int_{\frac{1}{2}\delta - u_2}^{\frac{1}{2}\delta - u_1} z(t)\,dt \doteq I_1 - I_2 + \tfrac{1}{2}\delta(z_1 - z_2).$$

Similarly,

$$p_3 \doteq \tfrac{1}{2} - I_1 + \tfrac{1}{2}\delta(z_0 - z_1), \qquad p_4 \doteq \tfrac{1}{2} - I_1 - \tfrac{1}{2}\delta(z_0 - z_1),$$

$$p_5 \doteq I_1 - I_2 - \tfrac{1}{2}\delta(z_1 - z_2), \qquad p_6 \doteq I_2 - \tfrac{1}{2}\delta z_2\,,$$

$$w_1 \doteq \frac{\delta z_2}{I_2}, \qquad w_2 \doteq \frac{\delta(z_1 - z_2)}{I_1 - I_2}, \qquad w_3 \doteq \frac{\delta(z_0 - z_1)}{\frac{1}{2} - I_1}.$$

Substituting these values in $E(x \mid A)$ and $V(x)$, retaining only the terms of order δ^2, we find

$$E(x \mid A) = \delta^2\left\{\frac{z_2^2}{I_2} + \frac{(z_1 - z_2)^2}{I_1 - I_2} + \frac{(z_0 - z_1)^2}{\frac{1}{2} - I_1}\right\}. \tag{6.3}$$

$$V(x) = 2E(x \mid A) = 2\,\delta^2 f(u_1\,, u_2), \tag{6.4}$$

writing $f(u_1\,, u_2)$ for the expression in brackets in (6.3).

We now reintroduce the subscript v in order to sum over the k variates. From inequality (6.2)′ the probability of misclassifying a specimen from A is the probability that the random variable $\sum (x_v \mid A)$ is negative. From (6.3) and (6.4) this variable is approximately normally distributed with

$$\text{Mean} = \sum_{v=1}^{k} \delta_v^2 f(u_{1v}\,, u_{2v}) : \ \text{Variance} = 2\sum_{v=1}^{k} \delta_v^2 f(u_{1v}\,, u_{2v}).$$

Hence the probability of misclassifying an A is approximately the probability that a normal deviate exceeds

$$\sqrt{\sum \delta_v^2 f(u_{1v}\,, u_{2v})/2}. \tag{6.5}$$

In order to minimize the probability we must maximize (6.5). Since $f(u_{1v}, u_{2v})$ does not depend on δ_v, expression (6.5) shows that the maximum is attained by substituting the *same* values of u_{1v}, u_{2v} in every term, i.e. the values which maximize $f(u_1, u_2)$. This reduces (6.5) to

$$\sqrt{\text{max. } f(u_1, u_2)} \sqrt{\sum \delta_v^2/2}. \qquad (6.6)$$

With the untransformed normal variate it was shown earlier that the best rule gives a probability of misclassification equal to the probability that a normal deviate exceeds $\frac{1}{2}\sqrt{\sum \delta_v^2}$. Comparing this result with (6.6), the relative discriminating power of the best qualitative to the best continuous rule is

$$2 \text{ max. } f(u_1, u_2) = 2 \text{ max. } \left\{ \frac{z_2^2}{I_2} + \frac{(z_1 - z_2)^2}{I_1 - I_2} + \frac{(z_0 - z_1)^2}{\frac{1}{2} - I_1} \right\}. \qquad (6.7)$$

This result holds in the sense that if a randomly chosen fraction 2 max. $f(u_1, u_2)$ of the original normal variates is retained, discarding the rest, the best continuous rule becomes equivalent to the best qualitative rule based on all the variates.

The result for five states is obtained by ignoring the distinction between the states α and β, calling the combined state D (doubtful). The effect is that the last term on the right of (6.7) disappears. For four states we let $u_1 = 0$, the states becoming A, a, b, B. For three states, the states a and b are combined into a D region. Finally, for two states we put $u_2 = 0$.

The form of the expression $2f(u_1, u_2)$, the values of u_1, u_2 giving the best partition of the normal curve, and the relative discriminating power of the qualitative variates appear in Table 7 for 2, 3, 4, 5 and 6 states.

Since the maxima of $f(u_1, u_2)$ are flat, the best values of u_1 and u_2 are given only to one decimal. For two states, the relative power is $2/\pi$, which will be familiar as the asymptotic power of the sign test. Use of five or six states retains over 90 percent of the power.

The function $2f(u_1, u_2)$ has already appeared in two other problems involving the replacement of normal curve methods by less efficient methods. Ogawa [1951] obtained this function as the efficiency of estimation of the population mean from suitably chosen order statistics in large samples. D. R. Cox [1957] found the same expression for the relative amount of information retained by grouping the normal curve when the group boundaries are chosen to minimize the quantity $E\{x - \mu(x)\}^2/\sigma^2$, where $\mu(x)$ is the mean of the group to which x is assigned. So far as we can see, the three problems are mathematically

TABLE 7

THE ASYMPTOTIC RELATIVE DISCRIMINATING POWER OF QUALITATIVE TO CONTINUOUS NORMAL VARIATES

No. of states	$2f(u_1, u_2)$	Best values of u_1	 u_2	Relative power
2	$2/\pi$	—	—	0.636
3	$2z_2^2/I_2$	—	0.6	0.810
4	$2\left\{\frac{z_2^2}{I_2} + \frac{(z_0 - z_2)^2}{\frac{1}{2} - I_2}\right\}$	—	1.0	0.882
5	$2\left\{\frac{z_2^2}{I_2} + \frac{(z_1 - z_2)^2}{I_1 - I_2}\right\}$	0.4	1.2	0.920
6	$2\left\{\frac{z_2^2}{I_2} + \frac{(z_1 - z_2)^2}{I_1 - I_2} + \frac{(z_0 - z_1)^2}{\frac{1}{2} - I_1}\right\}$	0.7	1.4	0.942

different, and the identity of results, as Cox has pointed out, seems to depend on particular properties of the normal curve.

7. COMPARISONS FOR SMALL NUMBERS OF VARIATES

As a check on the utility of the asymptotic results for practical situations, the best points of partition and the resulting probabilities of misclassification were computed directly for small numbers of variates. For simplicity the distance apart δ was assumed to be the same for all variates.

Having calculated the probability of misclassification for a given δ and number of variates k, we can compute the number k' of continuous normal variates that are needed to give the same probability of misclassification. For these values of δ and k, the ratio k'/k gives the relative discriminating power of the qualitative to the continuous normal variates. Unlike the asymptotic case, this ratio is of course dependent on the values of δ and k. For fixed k, however, the ratio turns out to be almost constant to two decimal places over the range of probabilities of misclassification that are of practical interest.

In the top half of Table 8, the relative discriminatory power with two variates is compared with the asymptotic power from Table 7

TABLE 8

Relative Discriminating Power of Qualitative to Continuous Normal Variates for Small Numbers of Variates

2 states		3 states		4 states		5 states	
$k = 2$	$k = \infty$	$k = 2$	$k = \infty$	$k = 2$	$k = \infty$	$k = 2$	$k = \infty$
.50	.64	.74	.81	.84	.88	.89	.92

2 states							3 states				
k							k				
2	3	4	5	6	7	∞	2	3	4	5	∞
.50	.74	.56	.70	.58	.68	.64	.74	.76	.77	.78	.81

for 2, 3, 4 and 5 states. The "small-sample" powers are all lower than the asymptotic values.

The lower half of Table 8 shows the results with two states and from 2 to 7 variates. The results have one curious feature. With an odd number of variates the relative powers are higher than the asymptotic values. The explanation may be that none of the decisions requires the toss of a coin when the number of variates is odd. In line with this result is the fact that with two states and only one variate ($k = 1$) the classification rule is the same as that obtained from the original normal variate, so that the relative power is unity. With 3 states (Table 8) this peculiarity does not appear, the relative powers increasing steadily with k towards the asymptotic value.

The optimum values of u_1 and u_2 for small numbers of variates were consistently close to the asymptotic optima. This result held throughout a wide range of values of $\delta \sqrt{k}$ from 0.5 (corresponding to a probability of misclassification of over 40 percent) to 4. Since, moreover, the optima are flat, the asymptotic values of u_1 and u_2 may be used safely with only two variates.

A few comparisons have been made between the qualitative and continuous discriminators for the case of two correlated variates of equal discriminating power. Suppose that the continuous scales are arranged so that the variates have standard deviations unity, with means zero in population A and δ in population B. If they are positively correlated, as appears to be the more common situation in applications, results obtained for 2 and 3 states indicate that the qualitative discriminator has higher relative power than when the variates are independent. The reverse seems to hold with negative correlation. Consequently our results for independent variates cannot be assumed to be valid for correlated variates. More investigation is needed.

In order to transform a continuous into a qualitative variate in an application we compute the means $\bar{y}_A$, $\bar{y}_B$ of the continuous variate in the initial samples from the two universes and the pooled within-universes standard deviation s. Let $\bar{y} = (\bar{y}_A + \bar{y}_B)/2$. With, for instance, five states the best asymptotic values of u_1 , u_2 are $u_1 = 0.4$, $u_2 = 1.2$. The states are constructed as follows.

Value of y	State
less than $\bar{y} - 1.2s$	A
from $\bar{y} - 1.2s$ to $\bar{y} - 0.4s$	a
from $\bar{y} - 0.4s$ to $\bar{y} + 0.4s$	D
from $\bar{y} + 0.4s$ to $\bar{y} + 1.2s$	b
greater than $\bar{y} + 1.2s$	B

8. ESTIMATION OF THE RELATIVE FREQUENCY WITH WHICH SPECIMENS COME FROM THE TWO UNIVERSES

If specimens present themselves with unequal frequencies π, π' from the two universes, the optimum rule is to assign a specimen to U if $\pi p_i > \pi' p_i'$. This rule requires an initial estimate or guess about the value of π. With qualitative variates the accuracy of this figure will often be not critical, because the same optimum rule holds over a range of values of π, although the range becomes shorter as the number of states increases. Table 9 gives an illustration for 4 states. The horizontal line marks the boundary of the best decision region.

TABLE 9

THE OPTIMUM RULE FOR DIFFERENT VALUES OF π

	Range of π .011 − .4		Range of π .4 − .64		Range of π .64 − .99	
State	p_i	p_i'	p_i	p_i'	p_i	p_i'
1	.92	.01	.92	.01	.92	.01
2	.06	.04	.06	.04	.06	.04
3	.04	.07	.04	.07	.04	.07
4	.00	.88	.00	.88	.00	.88

If we guessed $\pi = \pi' = 0.5$, we would use the middle classification rule in Table 9. This remains the best rule so long as π lies between 0.4 and 0.64. Even with π outside these limits, this rule may be close to the best. For instance, if π is 0.7, the average frequency of misclassification using our rule is (.7)(.04) + (.3)(.05), or .043, as against (.3)(.12), or .036 with the optimum rule. With $\pi = 0.8$ the corresponding figures are .042 and .024, a more serious difference.

When m specimens have been classified, the data give the numbers of specimens r_i found in each state, where $\sum r_i = m$. If the p_i, p_i' are known, the r_i follow a multinomial distribution with probability

$$P_i = \pi p_i + \pi' p_i'$$

in the ith state. The log likelihood is

$$L = \sum_1^s r_i \log(\pi p_i + \pi' p_i'), \qquad \frac{\partial L}{\partial \pi} = \sum \frac{r_i(p_i - p_i')}{\pi p_i + \pi' p_i'} \tag{8.1}$$

$$E\,\frac{\partial^2 L}{\partial \pi^2} = -E\left\{\sum \frac{r_i(p_i - p_i')^2}{(\pi p_i + \pi' p_i')^2}\right\} = -m \sum \frac{(p_i - p_i')^2}{\pi p_i - \pi' p_i'},$$

so that the estimated variance of $\hat{\pi}$ is

$$V(\hat{\pi}) = 1 \Big/ m \sum \frac{(p_i - p_i')^2}{\hat{\pi} p_i + \hat{\pi}' p_i'}. \tag{8.2}$$

The maximum likelihood estimate $\hat{\pi}$ obtained by setting $\partial L/\partial \pi = 0$ in (8.1) can be found by trial, although this is tedious if the states are numerous. One way of obtaining a first trial value of π is to note that each state furnishes an unbiased estimate. Since

$$E(r_i) = mP_i = m(\pi p_i + \pi' p_i'),$$

the estimate $\hat{\pi}_i$ from the ith state is

$$\hat{\pi}_i = [(r_i/m) - p_i']/(p_i - p_i') : V(\hat{\pi}_i) = P_i Q_i / m(p_i - p_i')^2.$$

These estimates are likely to differ markedly in precision. A fairly good method for a first trial $\hat{\pi}$ is to combine the states for which $p_i > p_i'$, making a single estimate from these states. Table 10 shows a numerical example for a sample of 100 specimens.

From the first two states a trial value is computed as

$$\hat{\pi}_1 = \frac{0.67 - 0.2}{0.6} = 0.78.$$

The values of $r_i(p_i - p_i')$, P_i and $r_i(p_i - p_i')/P_i$ are next computed (Table 10). The total of the latter values gives $\partial L/\partial \pi$ as -5.2. Since

TABLE 10

Example of the M. L. Estimation of π (Computed for $\hat{\pi} = .78$)

p_i	p_i'	r_i	$r_i(p_i - p_i')$	P_i	$r_i(p_i - p_i')/P_i$
.5	.1	37	14.8	.412	35.9
.3	.1	30	6.0	.256	23.4
.1	.2	9	−0.9	.122	−7.4
.1	.6	24	−12.0	.210	−57.1
		100			−5.2

$\partial L/\partial\pi$ is a decreasing function of π the second trial must be lower. The estimate $\hat{\pi}_2 = 0.75$ gave $+0.5$ for the first derivative. From (8.2) the estimated standard error of $\hat{\pi}_2$ was found to be ± 0.075.

Examination of the formula for the standard error shows that the estimate $\hat{\pi}$ cannot be expected to be precise: the standard error is always larger than the value $\sqrt{\pi(1 - \pi)/m}$ that would apply to a binomial estimate of π. This suggests that it will be worthwhile to estimate π only after a substantial number of specimens have been classified.

9. SUMMARY

This paper deals with the problem of assigning specimens to one of two or more universes when the measurements on each specimen are qualitative, each taking a small number of states. After presenting the optimum rule for classifying the specimens, three problems are considered.

The construction of the rule requires initial data on a number of specimens known to be classified correctly. Standard classification theory assumes that these initial samples are infinite in size, although in practice they may be only moderate. The principal effects of the finite sizes of the initial samples are that the probability of misclassification of the rule derived from them is underestimated and that this rule may be inferior to the theoretical optimum rule that we could construct if we had infinite samples. Methods are proposed for obtaining reasonably unbiased estimates of the performance of rules derived from finite samples and for estimating the difference between the actual and the theoretical optimum probability of misclassification. It appears that initial samples of size 50 from each of two universes should be adequate if there are not more than 8 multivariate states. With greater numbers of states, larger sample sizes are needed to ensure that the actual rule will be almost as good as the theoretical optimum.

If most of the variates are qualitative but a few are continuous, one possibility is to transform the continuous variates into qualitative ones, particularly since classification is easier with qualitative than with continuous variates. Asymptotic results are obtained for the best points of partition and the probabilities of misclassification when a large number of independent normal variates are partitioned to form qualitative variates. For qualitative variates with 2, 3, 4, 5 and 6 states the relative efficiencies are 64, 81, 88, 92 and 94 percent respectively. Computations for small numbers of variates show that the asymptotic points of partition remain satisfactory although the relative efficiencies are in general lower.

The optimum rule depends on the relative frequencies with which specimens to be classified present themselves from different universes. Initial estimates of these frequencies must be made in order to set up the rule. With two universes, maximum likelihood estimates of the frequencies from the data for specimens that have been classified by the rule are given. These estimates enable the rule to be improved if the initial estimates differ from the frequencies that apply when the rule is being used.

REFERENCES

Anderson, T. W. [1958]. *Introduction to multivariate statistical analysis.* John Wiley & Sons, New York, 142–7.

Cox, D. R. [1957]. Note on grouping. *Jour. Amer. Stat. Assn.* 52, 543–7.

Cox, D. R. and Brandwood, L. [1959]. On a discriminatory problem connected with the works of Plato. *Jour. Roy. Stat. Soc. B, 21*, 195–200.

Linhart, H. [1959]. Techniques for discriminant analysis with discrete variables. *Metrika 2*, 138–49.

Ogawa, J. [1951]. Contributions to the theory of systematic statistics. *Osaka Math. Jour. 4*, 175–213.

Rao, C. R. [1952]. *Advanced statistical methods in biometric research.* John Wiley & Sons, New York, Chapter 8.

Stokes, D. E., Campbell, A. and Miller, M. E. [1958]. Components of electoral decision. *Amer. Pol. Sci. Rev. 52*, 367–87.

THE STANDARDISATION OF TUBERCULIN HYPERSENSITIVITY

M. Stone
British Medical Research Council Applied Psychology Research Unit

R. A. Bruce
Chest Clinic, Cambridge, England

INTRODUCTION AND SUMMARY

Tuberculin hypersensitivity results from stimulation of the reticulo-endothelial system by tubercle bacilli and is due to the production of sensitised lymphocytes, that is, lymphocytes carrying antibody specific towards tuberculin. The presence of this antibody is detected by the Mantoux Test in which fluid containing tuberculin is injected into the superficial layers of the skin. This intradermal injection forms a bleb which quickly spreads outwards and is absorbed, depositing the tuberculin in the superficial layers. After deposition the tuberculin either escapes into the circulatory system or is reacted upon in situ by the antibody resulting in a final distribution of "reaction products" about the point of injection. It is postulated that induration or hardening of the skin will occur wherever the areal density of these products exceeds a certain level and that this level actually obtains at the edge of induration. In other words the areal density of tuberculin decreases from the point of injection and induration occurs wherever it exceeds a certain minimum level, the Minimum Tuberculin Concentration (c.f. Miles [1949]).

In this paper a mathematical model is presented by which the antibody-level may be calculated from a knowledge of the dose and concentration of tuberculin used and the resultant area of induration, provided this is non-zero. A statistical model of the multi-hit type deals with the antibody-level induced by different amounts of bacillary stimulation and thereby provides a measure of the virulence of the particular tubercle bacillus. The models can be applied to the results of skin-tests for any antibody by means of an appropriate antigen but for convenience they will be explained in the context of data from the Mantoux Test.

R. A. Fisher [1949] made a statistical analysis of the results of intradermal injections in cows in order to standardise different tuber-

TABLE 1

Effect of Tuberculin Dose in Constant Volume on Diameter of Induration

No. of subjects	Diameter mm		Areal Difference
	10 T.U.	1 T.U.	
Group I			
2	11.0	2.0	117
8	12.0	4.0	128
11	12.5	5.0	131
1	12.5	5.5	126
10	12.5	6.0	120
5	13.0	6.5	127
8	13.0	7.0	120
2	13.0	7.5	113
4	13.5	7.0	133
1	13.5	7.5	126
11	13.5	8.0	118
2	14.0	7.0	147
2	14.0	8.0	132
1	14.0	8.5	124
3	14.0	9.0	115
1	14.5	8.5	138
15	14.5	9.0	129
1	14.5	10.0	110
4	15.0	9.5	135
17	15.0	10.0	125
3	15.5	10.0	140
1	15.5	10.5	130
2	15.5	11.0	119
2	16.0	10.5	146
2	16.0	11.0	135
2	16.0	12.0	112
1	16.5	11.5	140
12	16.5	12.0	128
1	17.0	12.0	145
1	18.0	13.5	142
10	18.0	14.0	128
2	18.5	15.0	117
2	19.5	16.0	124
4	20.0	16.0	144
2	20.0	16.5	128
1	20.5	16.5	148
1	22.0	19.0	123
2	23.0	20.0	129
1	27.0	25.0	104

TABLE 1—(*Continued*)

No. of subjects	Diameter mm		Areal Difference
	10 T.U.	1 T.U.	
Group II			
10	12.0	4.5	124
11	12.5	6.0	120
5	14.5	9.0	129
5	15.0	10.0	125
3	20.0	16.0	144
6	20.0	16.5	128

culins. He concluded "the precision of such (tuberculin) readings regarded as biological assay seems to have been much underrated." Our work shows that a similar precision can be revealed by statistical and mathematical analysis for observations on human subjects and guinea-pigs.

EXPERIMENTAL MATERIALS AND METHODS

Tests were made on 260 human subjects and 10 guinea-pigs. The tuberculin was PPD (human) from the Ministry of Agriculture and Fisheries Institute, Weybridge. The B. C. G. was the liquid form from the Danish Serum Institute and the virulent tubercle bacillus was the strain Myc. Tub. H37Rv from the National Institute of Type Cultures. The injection-technique was carefully designed to reduce the error variance of the volumes injected. The diameters of induration were measured to 0.5 mm. at 72 hours after injection.

RESULTS AND THEORY

1. *The relation between induration-area and dose of tuberculin in constant volume*

About 250 subjects were given intradermal injections of 1 and 10 tuberculin units (T. U.'s) in 0.1 ml. Table 1 gives the results of those 161 subjects (Group I) who gave readable non-zero reactions to both doses. Another 40 subjects (Group II), known to be *negative reactors,* were BCG-vaccinated and then tested with 10 and 100 T. U.'s in 0.1 ml. Three levels of BCG-dose were used which explains the bunching of observations in this group.

The square of the diameter of induration D will be called the induration-area, written as D_1^2 , D_{10}^2 , D_{100}^2 for the three tuberculin doses.

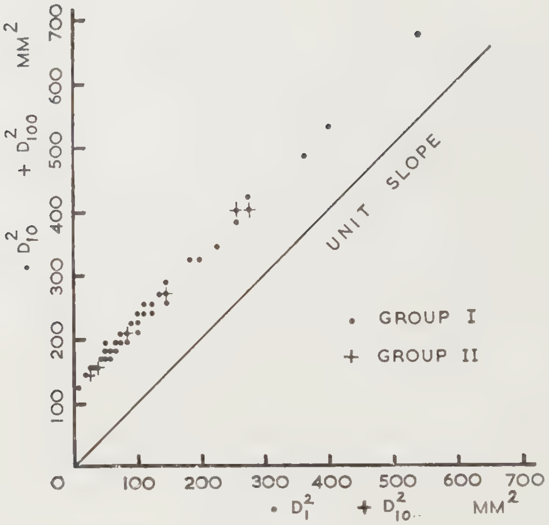


FIGURE 1

The Areas of the Paired Tuberculin Reactions of Groups I and II

The areas are plotted in Figure 1. The areal differences A (either $D_{10}^2 - D_1^2$ or $D_{100}^2 - D_{10}^2$) are given in the final column of Table 1. There is no discernible regularity in the A's of either group and 95 percent confidence limits for the difference of their means are -1.4, 3.8. Pooled, their standard deviation of 7.5 is consistent with constancy and an observational error of standard deviation 0.21 mm. For a representative areal difference is derived from reactions of sizes D' and D with means 15 and 10 mms. respectively. Then

$$\delta A \cong 30\ \delta D' - 20\ \delta D$$

$$\text{s.d. } A \cong (900 \text{ var } \delta D' + 400 \text{ var } \delta D)^{1/2} = 7.5.$$

These results suggest that the relation between the induration-area and the dose d of tuberculin in T. U.'s is

$$D^2 = \alpha^* \log d + \beta \tag{1}$$

where α^* is a constant for all 201 subjects and β is a personal constant containing all the information in the reaction relevant to the subject's sensitivity. The least-squares estimate of α^* is 126.7 (s. d. 0.5). For a given subject the critical dose of tuberculin which will just produce induration is obtained by putting $D = 0$ in equation (1).

Additional confirmation of the equation was provided by testing four subjects with a range of doses in 0.1 ml. The results are given

TABLE 2

The Observed and Predicted Induration Areas at Constant Volume

Subjects	Dose of PPD in T.U. (d)						
	10	5	2.5	1	0.5	0.25	0.1
1	256	225	196	—	100	—	6
		(218)	(180)		(91)		(3)
2	144	110	72	20	0	0	0
		(106)	(68)	(17)	(0)	(0)	(0)
3	144	110	72	16	0	0	0
		(106)	(68)	(17)	(0)	(0)	(0)
4	110	—	36	—	—	0	0
			(34)			(0)	(0)

in Table 2 where the bracketed figures are the areas predicted from equation (1) with $\alpha^* = 126.7$, using the observed areas at 10 T. U. to estimate β.

2. *Effect of change of volume of injected fluid on (i) induration-area (ii) area covered by fluid.*

16 subjects were each given four simultaneous intradermal tuberculin injections, 10 and 1 T. U. in 0.1 and 0.2 ml., and two intradermal injections of Evans blue in volumes of 0.1 and 0.2 ml. The results are given in Table 3 where F is the diameter to which the dyed fluid spread after two hours. For the tuberculin results, at each volume the areal difference A from 1 to 10 T. U. is calculated for the cases when both areas are non-zero. The average of the A's at 0.1 ml. is 133.2 which is not significantly different from 126.7. Half the average at 0.2 ml. is 121.8 which is only 4.9 less than 126.7. The suggestion is that the rate of change of area with log-dose is proportional to the volume of fluid V. (In fact, for all but one subject the A at 0.2 ml. is less than twice the A at 0.1 ml., so that the relation is probably not exact.) Equation (1) can be extended to give

$$D^2 = \bar{\alpha} V \log d + \beta(V). \qquad (2)$$

Comparison with equation (1) shows that $\beta(0.1) = \beta$ and $0.1\ \bar{\alpha} = \alpha^*$, so that $\bar{\alpha}$ can be estimated as 1267. For the dye-injections, the results suggest that F^2/V is independent of V or that

$$F^2 = \kappa V. \qquad (3)$$

TABLE 3

The Effect of Volume of Injected Fluid on the Induration-Area and the Area Covered by the Fluid

Vol. injected fluid, V	0.1 ml.					0.2 ml.				
	Tuberculin			Dye		Tuberculin			Dye	
Subject	10 T.U. Diam.	1 T.U. Diam.	Areal difference	Diam. F	F^2/V or κ	10 T.U. Diam.	1 T.U. Diam.	Areal difference	Diam. F	F^2/V or κ
1	4.0	0.0		20	4000	0.0	0.0		28	3920
2	5.0	0.0		13	1690	0.0	0.0		18	1620
3	7.0	0.0		13	1690	5.0	0.0		18	1620
4	12.5	5.0	131	13	1690	15.5	0.0		18	1620
5	15.0	10.0	125	20	4000	18.5	11.0	221	28	3920
6	15.0	10.0	125	20	4000	19.0	11.0	240	28	3920
7	15.0	10.0	125	20	4000	19.5	11.5	248	28	3920
8	16.5	12.0	128	20	4000	21.5	15.0	237	28	3920
9	16.5	12.0	128	16	2560	22.0	15.0	259	22	2420
10	17.0	12.0	145	20	4000	22.5	16.0	250	28	3920
11	17.0	12.0	145	20	4000	23.0	16.5	257	28	3920
12	18.0	14.0	128	20	4000	24.0	18.0	252	28	3920
13	19.0	15.0	136	20	4000	25.0	19.0	264	28	3920
14	20.0	16.5	128	29	8410	27.0	22.0	245	41	8405
15	23.0	20.0	129	29	8410	31.0	27.0	232	41	8405
16	28.0	25.0	159	29	8410	38.0	35.0	219	41	8405

Model for the deposition of tuberculin

By requiring that $d \exp(-D^2/\bar{\alpha}V \log_{10} e)$ be constant for constant V, equation (2) suggests that the fluid spreads and is absorbed *as if*, before absorption, it assumed the shape of a circular-normal distribution truncated at diameter F and with a variance s^2 proportional to V.

For such a distribution the areal concentration of fluid $f(r)$ at radius r would be

$$f(r) = [V \exp(-\tfrac{1}{2}r^2/s^2)]/2\pi s^2[1 - \exp(-F^2/8s^2)]$$

for $r < \tfrac{1}{2}F$. If δV is the extra fluid conceptually needed to extend the circular-normal distribution to infinity then

$$V/(V + \delta V) = 1 - \exp(-F^2/8s^2) = \mu$$

say. If $8s^2 = \alpha V$ where α is a constant, then use of equation (3) gives

$$\mu = 1 - \exp(-\kappa/\alpha) \tag{4}$$

so that μ is also a constant. If $t(r)$ is the areal density of deposited tuberculin at radius r in T. U.'s/mm^2, then

$$t(r) = 4\, d \exp(-4r^2/\alpha V)/\pi\alpha\mu V. \tag{5}$$

Denoting the Minimum Tuberculin Concentration by τ, this will obtain at the edge of induration provided $D < F$; so substituting $r = \tfrac{1}{2}D$ and $t(\tfrac{1}{2}D) = \tau$ in equation (5), we obtain

$$D^2 = \alpha V \log_e (4\, d/\pi\alpha\mu\tau V). \tag{6}$$

When $D = F$, the Minimum Tuberculin Concentration may be exceeded at the edge of induration and equation (6) will not be true. Reference to Table 3 shows that usually $D < F$.

Agreement with the data

Equation (2) agrees with equation (6) when $\bar{\alpha} = \alpha \log_e 10$ and $\beta(V) = \bar{\alpha}V \log (4 \log_e 10/\pi\bar{\alpha}\mu\tau V)$. By equation (4) we find that the four levels of κ represented in Table 3 have corresponding levels of $1 - \mu$ of $10^{-1.3}$, $10^{-2.0}$, $10^{-3.2}$ and $10^{-6.6}$. So for these subjects $\mu \cong 1$ and the relation becomes

$$D^2 = \bar{\alpha}V \log (4\, d \log_e 10/\pi\bar{\alpha}\tau V) \tag{7}$$

to a good approximation. (It is, however, conceivable that, if more subjects had been examined, smaller values of μ would have been found.) Equation (7) predicts the relationship between D and V when d is held constant. Table 4 gives the results of calculations of $\log \tau$ for the subjects of Table 3. For each subject, the model is supported by (i) the constancy of the log τ's calculated from non-zero diameters, (ii) the fact that for a zero diameter the log τ calculated by putting $D = 0$ in equation (7) is smaller. From equation (7) it follows that, when V is increased, D^2 will also increase but that a volume is ultimately reached above which D^2 will decrease. When $d = 10$ T. U.'s and log τ exceeds -1.238 or when $d = 1$ T. U. and log τ exceeds -2.238, we would expect the D^2 for 0.1 ml. to *exceed* that for 0.2 ml. The expectation is supported by subjects 1, 2, 3 and 4. For subjects 1 and 2 at 10 T. U.'s and subject 4 at 1 T. U., the effect of increasing V is to dilute the tuberculin so much that at no point in the entire area of the reaction is the Minimum Tuberculin Concentration reached and no induration develops. For subjects 5–16 this dilution effect is more than offset by the increased spread of the fluid which results in tuberculin being carried further from the site of injection at more than the necessary minimum concentration.

3. *Antibody level related to (i) Minimum Tuberculin Concentration and (ii) bacillary stimulation.*

(i) Let $t(T)$ be the areal density of tuberculin at a particular point at time T after deposition. The rate of escape of tuberculin into the circulatory system without reaction with antibody will be $[t'(T)]_E = -\epsilon t(T)$ where ϵ is a constant for a given subject (Kety [1949]). If λ is the antibody density, the rate of accumulation of reaction products will be $r'(T) = \lambda t(T)$ and the rate of reaction of tuberculin will be $[t'(T)]_R = -\omega r'(T)$ where ω is a constant. Therefore

$$t'(T) = [t'(T)]_E + [t'(T)]_R = -(\epsilon + \omega\lambda)t(T)$$

$$t(T) = t(0) \exp [-(\epsilon + \omega\lambda)T]$$

$$r(T) = \lambda t(0)\{1 - \exp [-(\epsilon + \omega\lambda)T]\}/(\epsilon + \omega\lambda)$$

$$r(\infty) = \lambda t(0)/(\epsilon + \omega\lambda).$$

If induration occurs wherever $r(\infty) > \rho$ then

$$\tau = \gamma/\lambda + \delta \tag{8}$$

where $\gamma = \epsilon\rho$ and $\delta = \omega\rho$.

(ii) *The multi-hit model.*

Injected (or otherwise acquired) bacilli will be distributed throughout the reticulo-endothelial tissues which consist of *lymphocyte-producing centres*. The bacilli stimulate these centres to pour sensitised lymphocytes (antibody) into the circulation. Suppose (a) the bacilli are distributed randomly among a much larger number of centres, (b) at least ν bacilli are required to sensitise a centre, (c) the resulting λ is proportional to the number of sensitised centres. If B bacilli are distributed among C centres and $p_r(B)$ and $s_r(B)$ are the mean proportions of centres with r and r or more bacilli respectively, then

$$Cs_r(B+1) - Cs_r(B) = p_{r-1}(B) \qquad r = 1, 2, \cdots$$

or approximately, when B is large,

$$Cs_r'(B) = s_{r-1}(B) - s_r(B).$$

Therefore

$$\left(C\frac{d}{dB} + 1\right)^r s_r(B) = s_0 = 1$$

the solution of which is

$$\begin{aligned} s_r(B) &= 1 - e^{-B/C}[1 + (B/C) + \cdots + \{1/(r-1)!\}(B/C)^{r-1}] \\ &= e^{-B/C}[(1/r!)(B/C)^r + \{1/(r+1)!\}(B/C)^{r+1} + \cdots]. \end{aligned}$$

TABLE 4

Log Minimum Tuberculin Concentration (τ)

Injection-volume	0.1 ml.		0.2 ml.		Average log τ	$\bar{\kappa}$
Dose of PPD	10 T.U.	1 T.U.	10 T.U.	1 T.U.		
Subject						
1	−0.76	(−1.64)	(−0.94)	(−1.94)	−0.76	3960
2	−0.83	(−1.64)	(−0.94)	(−1.94)	−0.83	1655
3	−1.02	(−1.64)	−1.04	(−1.94)	−1.03	1655
4	−1.87	−1.83	−1.88	(−1.94)	−1.86	1655
5	−2.41	−2.42	−2.29	−2.41	−2.38	3960
6	−2.41	−2.42	−2.36	−2.41	−2.38	3960
7	−2.41	−2.42	−2.44	−2.46	−2.43	3960
8	−2.78	−2.77	−2.76	−2.82	−2.78	3960
9	−2.78	−2.77	−2.85	−2.82	−2.80	2490
10	−2.92	−2.77	−2.93	−2.95	−2.89	3960
11	−2.92	−2.77	−3.02	−3.01	−2.93	3960
12	−3.19	−3.18	−3.21	−3.22	−3.20	3960
13	−3.48	−3.41	−3.40	−3.36	−3.41	3960
14	−3.79	−3.78	−3.81	−3.85	−3.81	8408
15	−4.81	−4.79	−4.73	−4.81	−4.78	8408
16	−6.82	−6.57	−6.64	−6.77	−6.70	8408

If B/C is small, we therefore expect $\lambda \propto B^{\nu}$ or

$$\log \lambda = \nu \log B + \eta \tag{9}$$

where η is a constant. The integer ν provides the measure of bacillary virulence.

Agreement with data.

(i) *Human Subjects.* 39 female subjects, whose preliminary skin-tests showed no induration, were vaccinated with a range of B. C. G. doses and after six weeks were tested with 100 T. U.'s in 0.1 ml. The log τ's calculated from equation (7) are given in Table 5. Each subject's B. C. G. dose establishes a level of λ. Table 5 shows that the variance

TABLE 5

Effect of B.C.G. Dose on Size of Induration

No. of subjects	Vol. B.C.G. in ml.	Induration diameter (mms)	log τ	Average log τ
2	0.300	32.0	−7.72	−7.80
1	0.300	32.5	−7.97	
2	0.250	30.0	−6.74	−6.90
1	0.250	31.0	−7.22	
1	0.245	28.0	−5.82	−6.16
1	0.245	29.5	−6.50	
1	0.155	25.5	−4.77	−4.77
1	0.150	24.0	−4.18	−4.47
3	0.150	25.0	−4.57	
1	0.140	24.0	−4.18	−4.18
1	0.130	23.0	−3.81	−3.81
2	0.120	22.0	−3.46	
1	0.120	22.5	−3.63	−3.59
1	0.120	23.0	−3.81	
1	0.100	20.0	−2.79	−2.79
1	0.080	16.0	−1.66	−1.79
1	0.080	17.0	−1.92	
1	0.075	14.5	−1.30	
3	0.075	15.0	−1.41	−1.41
1	0.075	15.5	−1.53	
2	0.070	14.0	−1.18	−1.30
1	0.070	15.5	−1.53	
1	0.065	12.0	−0.77	−0.77
3	0.060	10.0	−0.43	
2	0.060	10.5	−0.51	−0.51
3	0.060	11.0	−0.59	

of Minimum Tuberculin Concentrations within dose-levels is much smaller than the between-dose variance. This implies that the constants γ and δ differ only slightly, if at all, among these subjects. If δ is constant from subject to subject, it is clear from equation (8) that it is less than any τ occurring. Therefore $\delta < 10^{-7.97}$. The average of

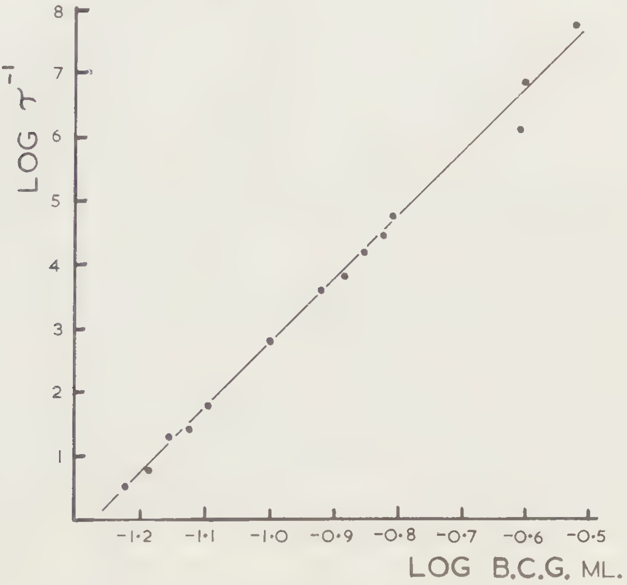


FIGURE 2

THE EFFECT OF B.C.G. DOSE ON THE MINIMUM TUBERCULIN CONCENTRATION

log τ^{-1} is plotted against the log B. C. G.-dose in Figure 2. The least-squares straight-line has the equation

$$\log \tau^{-1} = 10.1 \log B.C.G. + 12.9. \tag{10}$$

(95 percent confidence limits for the slope are 9.8, 10.4. A significantly better fit is possible by allowing a quadratic term but the improvement is only slight.) It is reasonable to extrapolate and state that higher B. C. G. doses would result in much lower τ's than 10^{-8} T. U.'s/(mm)^2. So for all subjects we will put $\delta = 0$ in equation (8) and write

$$\tau = \gamma/\lambda. \tag{11}$$

Since $\gamma = \epsilon\rho$ and $\delta = \omega\rho$, the interpretation of $\delta \cong 0$ is that most of the tuberculin escapes without local reaction.

As stated above, the variance of γ from subject to subject must be quite small. This is somewhat surprising since γ is presumably related to skin-type. However, all 39 subjects were aged 18 or 19 so that perhaps γ is determined by age. The subjects of Table 4 throw some light on this. The relation of the *area of fluid coverage* constant κ with age is as follows:

$$\begin{aligned} \kappa &= 1655, \text{ ages over 65;} \\ \kappa &= 2490, \text{ age 55;} \\ \kappa &= 3960, \text{ ages between 30 and 45;} \\ \kappa &= 8408, \text{ ages under 30.} \end{aligned}$$

The negative correlation is statistically significant ($P = 0.05$). There is a (non-significant) negative correlation of κ and τ. Accepting both correlations as real, there would be two possibilities consistent with the theory: (i) γ is independent of age but both κ and λ decrease with age, the latter due to decreasing bacillary stimulation. In this case the determination of κ by dye-injection has no relevance to the interpretation of the Mantoux reaction. (ii) γ and κ are correlated, both being functions of skin-type. The data considered does not distinguish these two. One experimental approach would be to take *negative reactors* at different ages and induce possibly identical antibody levels by B. C. G. vaccination.

Equations (10) and (11) combine to give

$$\log \lambda = 10.1 \log B.C.G. + 12.9 + \log \gamma.$$

The confidence interval for the slope includes the value 10 so that the results are consistent with equation (9) when $\nu = 10$.

(ii) *Guinea-pigs.*

Each of five dose-levels of virulent Myc. Tub. bacilli was given by intraperitoneal injection to one of five pairs of guinea-pigs. The animals were tested with 100 T. U.'s in 0.1 ml. on the 10th, 18th, 27th and 35th days after injection. On one occasion each pair was also tested with 320 T. U.'s. The average diameters for each pair are given in Table 6 with the 320 T. U. results in brackets. The simultaneous diameters at 100 T. U. and 320 T. U. are consistent with equation (1) with the same value of α^* as for humans 126.7. For the differences $D_{320}^2 - D_{100}^2$ are 56, 56, 61, 66 and 69 with an average of 62. By equation (1) the difference should be 64. If equations (7) and (11) also hold, the antibody levels, $\log (\lambda/\gamma)$, can be calculated. These are plotted in Figure 3. For these animals the bacillary content is increasing with time. If the increase is geometric, the actual bacillary content

TABLE 6

THE INDURATION DIAMETERS PRODUCED BY DIFFERENT DOSES OF MYCOBACTERIUM TUBERCULOSIS (H 37 Rv) IN GUINEA-PIGS

Days after infection	Dose of infecting organisms; Log mgms. wet weight of bacilli				
	−1.796	−1.893	−1.990	−2.087	−2.184
10	9	8	5	2	0
18	10	9	7	5 (9)	0
27	11	10 (12.5)	8.5	7	5
35	12 (14.5)	11	10 (13)	8.5	7 (10.5)

on the Tth day will be $B_0 10^{kT}$ where B_0 is the initial dose and k is the growth-rate constant. If equation (9) holds, then

$$\log (\lambda/\gamma) = \nu \log B_0 + \nu k \log T + \text{const.}$$

The least-squares estimates are $\hat{\nu} = 2.05$, $\hat{k} = 0.0104$ and $\widehat{\text{const.}} = 3.72$. The lines drawn in Figure 3 are based on these estimates. In view of the small number of observations a slope of 2.05 is consistent with a true value of ν of 2.

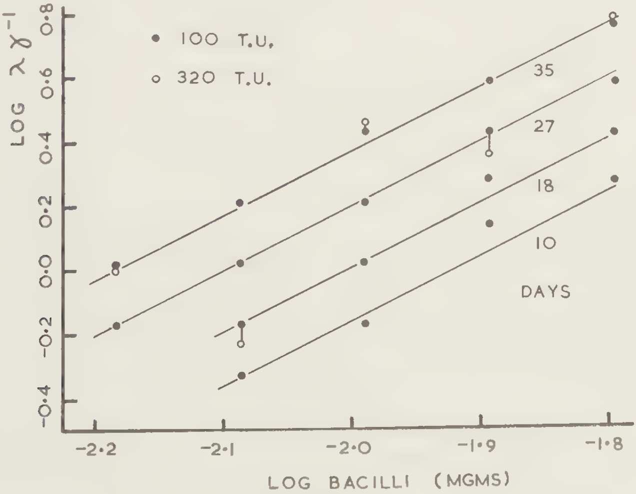


FIGURE 3

THE RELATION BETWEEN ANTIBODY LEVEL AND INFECTING DOSE OF MYC. TUB. (H 37 Rv) AT 10, 18, 27 AND 35 DAYS AFTER INJECTION

If the models are valid, it must be concluded that, whereas at least 10 B. C. G. bacilli are needed to activate a lymphocyte-producer in humans, only two virulent Myc. Tub. bacilli are required in guinea-pigs.

DISCUSSION

The conclusion of Section 1 of Results was that there exists a linear relation between the induration-area and the log-dose of tuberculin in constant volume. This contradicts the conclusion of Long and Miles [1950] with tuberculin and of Miles [1949] with diphtheria toxin that the linear relation is between *diameter* and log-dose. In Long and Miles [1950] it is stated that Wadley [1948, 1949] had found such a relation. However examination of Wadley's paper shows that the bulk of the work established a linear relation between log-dose and the *increase in thickness* of the induration produced by tuberculin injection in cattle. Miles [1949] rests the case for his linearity on visual inspection and analyses of variance. In his Table 1 there is a statistically significant departure from linearity but the associated F-value is small compared with that for the linear component. However, provided the range of reaction-diameters is not too large, this may also be expected to result from our equation (1). For example, with five doses at two-fold dilution intervals and the diameters ranging from 10 to 17.5 mms, the mean square for linearity would be several hundred times larger than the mean square for departures from linearity. In other words for this range the two theories are not clearly distinguishable. Miles observes, and it is apparent from his figures, that the departures from linearity tended to occur with small induration-diameters—"the curves dipping more steeply to the 3–5 mm values". He invokes a special argument in explanation of this. However this kind of departure is precisely what our equation (1) would lead us to expect.

In this paper an attempt has been made to present a validation of two models without considering whether they can be justified on more general grounds. A fuller discussion of their clinical and immunological implications will be published elsewhere.

Further experimentation is clearly needed to deal with such questions as the following:

(i) Are the constants or even the form of the dose-response relation of the intradermal test dependent on experimental technique?

(ii) Is the multi-hit model correct for bacilli other than BCG and the Mycobacterium Tuberculosis used?

(iii) Is the approach adopted useful for diseases other than tuberculosis?

ACKNOWLEDGEMENTS

We wish to thank the Medical Research Council and the East Anglian Regional Hospital Board for financial assistance. For experimental facilities we are grateful to Professor L. P. Pugh of the Cambridge Department of Veterinary Medicine, to Dr. S. H. Currie, the Staff and Employees of Pye Ltd., Cambridge, and to Dr. D. H. Clark, Medical Superintendent of Fulbourn Hospital. We are also grateful to referees for pointing out a numerical error and for suggesting improvements in presentation.

REFERENCES

Fisher, R. A [1949]. A biological assay of tuberculins. *Biometrics 5*, 300–316.

Kety, S. S. [1949]. Measurement of the regional circulation by the local clearance of radioactive sodium. *Amer. Heart J. 38*, 321.

Long, D. A. and Miles A. A. [1950]. Opposite actions of thyroid and adrenal hormones in allergic hypersensitivity. *Lancet* i, 492–5.

Miles A. A [1949]. The fixation of diphtheria toxin to skin tissue with special reference to the action of circulating antitoxin. *Brit. J. Exp. Path. 30*, 319–44.

Wadley, F. M. [1948]. Experimental design in comparison of allergens on cattle. *Biometrics 4*, 100–8.

Wadley F. M. [1949]. Use of biometric methods in comparison of acid-fast allergens. *Amer. Rev. Tub. 60*, 131–9.

OPTIMUM ESTIMATION OF GRADIENT DIRECTION IN STEEPEST ASCENT EXPERIMENTS

SAMUEL H. BROOKS AND M. RAY MICKEY
C-E-I-R, Inc.
Los Angeles, California, U.S.A.

Summary

In the steepest ascent method for seeking maxima (Box and Wilson), the procedure involves the progression to a region of high response by successive inferences as to the gradient direction in sets of trials en route. When there is no experimental error, and the response surface is well approximated by a plane in any locality, then it can be seen that an appropriately patterned set of trials, numbering just one more than the number of factors, would be required to completely characterize this plane and thus the gradient direction. Were there experimental error, however, the gradient direction would be correspondingly uncertain. One wonders whether to expect that more ground would be gained by having more trials in each set at the expense of having fewer sets. As this problem is formulated herein, there is a strong conclusion that the number of trials per set should be minimal.

I. *Introduction*

The procedure of steepest ascent (Box and Wilson, [1951]) consists of performing a sequence of sets of trials. Each set of trials is "centered" at a combination of factor levels, or point in factor space. An initial point is selected for the center of the first set of trials. Succeeding center points are determined by estimating the appropriate relative changes among the factor levels, i.e., the direction of the change, from the results of the set of trials just performed and scaling the increments of change such that the new center point is a fixed distance, the step length, away from the old. This procedure is followed until a near stationary region is attained.

This investigation is concerned with the allocation of effort to the trials in the initial stages of the search, where one is not likely to be close to a point of maximum response. It is assumed that the choice of scales for the factor levels has been decided upon, so that a distance is defined, and that a step length has been selected. The purpose of a set of trials is (and we here regard this as the sole purpose) then to supply the direction of the "step" for determining the next center point. Box and Wilson [1951] have shown that the appropriate direction is that of the gradient at the new point. Consequently the purpose of a set of trials is to provide an estimate of the gradient direction.

It is assumed that the response surface is well approximated by a plane in any locality. Then the gradient direction estimated as a result of a set of trials differs from that of the true gradient by some angle, say θ, when there is experimental error. The greater the number of trials in a set, the smaller θ is expected to be. Let us designate by S the size of step to be taken in the estimated gradient direction. As a result, $S \cos \theta$ is the component of this step which is in the direction of the true gradient. If $\cos \theta$ were near to one, the gradient would be well estimated and the step would be well taken. If $\cos \theta$ were near zero, there would be comparatively little change in the response to be expected. If $\cos \theta$ were negative, the step would be downhill. Thus, $S \cos \theta$ may be regarded as the improvement in position in factor space as a result of a set of trials.

Let us regard a trial as a unit of effort, so that for t trials in a set, the improvement per unit of effort is $(S \cos \theta)/t$. In planning a set of trials it is desirable to select t so as to maximize the expected value of the improvement per unit effort, i.e. $E[(S \cos \theta)/t]$. This is equivalent to selecting t so as to maximize $(1/t)\, E(\cos \theta)$, where $E(\cos \theta)$ is regarded as a function of t.

Our main result is that the maximum is achieved when t is one more than the number of factors; that is, it is not worthwhile to improve the precision of the estimate of the gradient direction at the expense of more trials. Since this result would be anticipated if the experimental variation could be ignored, it is of interest that it applies independently of the magnitude of the variability. The result is convenient from the point of view of application, since otherwise the appropriate value of t would depend upon the magnitude of the gradient and upon the experimental variability, the ratio of which would have to be estimated in advance of the set of trials. Consequently, the simplicity of the result enhances its value.

The result is developed from the assumption that the trial design yields normally distributed uncorrelated estimates of the gradient components which have the same variance. The magnitude of the common variance will depend upon the design used. Although the present result does not evaluate the relative efficiency of alternative designs, the above formulation of the problem does provide a basis for comparison by means of $E(\cos \theta)$. While we do not pursue this line of investigation here, some tabular values are given, so that alternative designs can be readily compared from the point of view of efficiency for maxima seeking.

The basic function $E(\cos\,\theta)$ is evaluated in Section II, and some numerical values are presented. The development is completed in Section III by applying the result of Section II to establish the main result of the paper. Some numerical values are presented of how the efficiency depends on the number of trials within a set.

Brief mention should be made of related aspects of the maximum seeking problem that our formulation has not considered. The problem of recognizing that a near stationary region has been reached may present difficulties in the case of large experimental error. Another aspect is the effect of the curvature of the response surface both on the selection of the design of a set of trials and the size of the region over which the design is to be applied and consequently on the estimation of the gradient direction. Finally, we note that the prior selection of scales of the factor levels and step length also affect the estimate of the gradient direction.

II. *Derivation of E(cos θ)*

Let us suppose that in the vicinity of a set of trials the relation between the response and the n factors is well represented by a plane:

$$R = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon,$$

where x_i is the level of the ith factor and ϵ is the experimental error,[1] $NID(0, \sigma_\epsilon^2)$.

Let us suppose, also, that as a result of a set of t trials it is possible to construct an estimate b_i of β_i such that the b_i are $NID(\beta_i\,, \sigma^2)$ and σ^2 is a function of t.

The angle between the true gradient β, having components β_i , and the estimated gradient b, having components b_i , is θ. Consequently:

$$\cos\,\theta = \sum b_i\beta_i/\sqrt{\sum b_i^2 \sum \beta_i^2} \quad \text{where} \quad 0 \leq \theta \leq \pi.$$

$E(\cos\,\theta)$ is to be found from the joint distribution of the magnitude and angle of error, θ of the random vector b. It is possible to orient the factor space so that, without loss of generality, the components of β may be expressed, in units of σ, as $\rho, 0, 0, \cdots, 0$. Under the same transformation the components of b are $y_1\,, y_2\,, \cdots, y_n$. The y_i are NID with

$$E(y_i) = \begin{cases} \rho & i = 1 \\ 0 & \text{Otherwise} \end{cases}$$

[1] NID $(0, \sigma_\epsilon^2)$ means normally and independently distributed with mean zero and variance σ_ϵ^2.

and each with unit variance. The magnitude of b, in units of σ, is r. The joint distribution of r and θ is found from the joint distribution of the y_i :

$$dF(y_i) = (2\pi)^{-n/2} e^{-1/2[(y_1-\rho)^2+y_2^2+\cdots+y_n^2]}\, dy_1\, dy_2 \cdots dy_n\,.$$

Noting that the quantity in the brackets can be expressed by the cosine law, this function, in terms of θ and r, becomes:

$$dF(r,\theta) = (2\pi)^{-n/2} e^{-1/2(r^2-2r\rho\cos\theta+\rho^2)} r\, d\theta\, dr\, dH,$$

where dH is the region over which r and θ are constant; that is, the surface "area" of an $(n-1)$-dimensional hypersphere of radius $r \sin \theta$. By integration over this region:

$$dF(r,\theta) = \frac{r^{n-1}(\sin\theta)^{n-2}}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)2^{(n-2)/2}}\, e^{-1/2(r^2-2r\rho\cos\theta+\rho^2)}\, dr\, d\theta.$$

This provides an expression for the expected value of $\cos\theta$:

$$E(\cos\theta) = \int_{\theta=0}^{\pi} \int_{r=0}^{\infty} \cos\theta\, dF(r,\theta).$$

Expanding $e^{r\rho\cos\theta}$ in powers of $r\rho\cos\theta$ leads to:

$$E(\cos\theta) = e^{-\rho^2/2} \sum_{j=0}^{\infty} \frac{\rho^j}{j!} \int_0^{\pi} \frac{(\cos\theta)^{j+1}(\sin\theta)^{n-2}}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)2^{(n-2)/2}}\, d\theta \int_0^{\infty} r^{n+j-1} e^{-r^2/2}\, dr.$$

Each of the even j terms in the above expression are zero since, for even j,

$$I \equiv \int_0^{\pi} (\cos\theta)^{j+1}(\sin\theta)^{n-2}\, d\theta = 0.$$

For odd j this I integral may be put in the form of the beta function by letting $X = \sin^2\theta$:

$$I = 2\int_0^{\pi/2} (\cos\theta)^{j+1}(\sin\theta)^{n-2}\, d\theta$$

$$= \int_0^1 (1-X)^{j/2} X^{(n-3)/2}\, dX = \frac{\Gamma\left(\frac{j+2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{j+n-1}{2}\right)}.$$

The r integral evaluated as a gamma function becomes:

$$2^{(n+j-2)/2}\Gamma\left(\frac{n+j}{2}\right).$$

TABLE 1

Expected Value of cos θ for n, the Number of Factors, and for ρ, the Magnitude of the True Gradient Relative to the Standard Error of an Estimated Component as Computed from the Exact Formula

Number of Factors, n	True gradient ρ (in units of the standard error of component estimation)							
	0.0	0.25	0.50	1	2	4	8	∞
2	0.0000	0.1554	0.3038	0.5572	0.8443	0.9669	0.9921	1.0000
4	0.0000	0.1769	0.2302	0.4342	0.7140	0.9109	0.9768	1.0000
8	0.0000	0.0854	0.1692	0.3266	0.5759	0.8250	0.9482	1.0000
16	0.0000	0.0614	0.1222	0.2395	0.4447	0.7097	0.8984	1.0000
32	0.0000	0.0438	0.0874	0.1729	0.3317	0.5773	0.8186	1.0000

Assembling these terms and letting $j = 2i + 1$, there results:

$$E(\cos\theta) = e^{-\rho^2/2} \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{2i+3}{2}\right) 2^{(2i+1)/2} \Gamma\left(\frac{n+1}{2} + i\right) \rho^{2i+1}}{(2i+1)!\, \Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n+2}{2} + i\right)}$$

$$= \frac{\rho}{\sqrt{2}} e^{-\rho^2/2} \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{n+1}{2} + i\right)\left(\frac{\rho^2}{2}\right)^i}{\Gamma\left(\frac{n+2}{2} + i\right) i!}.$$

It may be noted that the above expression is closely related to the confluent hypergeometric function.

An approximation to $E(\cos\theta)$ follows directly from the expression given for $\cos\theta$ by noting that

$$E(\sum b_i\beta_i) = \sum \beta_i^2 = \rho^2\sigma^2,$$

$$E(\sum b_i^2) = (n + \rho^2)\sigma^2.$$

By substituting expected values for the random variables one obtains:

$$E(\cos) \cong \rho/\sqrt{n + \rho^2}.$$

This approximation is in good agreement with the values in the table and illustrates the general nature of this function.[2]

III. *Optimum Number of Trials in a Set*

In the previous section it was stated that ρ is in units of σ and that σ is a function of t. $E(\cos\theta)$ is to be expressed as a function of t so that $(1/t)E(\cos\theta)$ can be examined for the t which maximizes this quantity; that is, the number of trials in a set which results in the greatest expected improvement per unit of effort.

Since the β_i are ordinarily estimated by means of linear functions of the observations, then σ^2 would ordinarily be a simple multiple of σ_ϵ^2. Were that set of trials replicated, say m times, then σ^2 would be the same multiple of σ_ϵ^2/m, which would be proportional to σ_ϵ^2/t. Under these circumstances σ^2 would be proportional to $1/t$. It is suggested that this would be a good approximation of the functional relation of σ^2 and t for more varied experimental designs than those which are replicated, as long as the trials would be made in the same locality. Consequently, since ρ^2 is in units of σ^2, ρ^2 is proportional to t. It is

[2] It can be shown that a lower bound for $E(\cos\theta)$ is $\rho/\sqrt{n + \rho^2 + 1}$.

convenient to work with a quantity which is proportional to t and therefore to ρ^2. Let T, k and $h_n(T)$ be defined by the following:

$$T \equiv kt \equiv \rho^2/2,$$

$$(1/t)E(\cos\theta) = h_n(T) \equiv ke^{-T} \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{n+1}{2} + i\right)T^{(i-1/2)}}{\Gamma\left(\frac{n+2}{2} + i\right)i!}.$$

As a result of differentiating this with respect to T:

$$\frac{dh_n(T)}{dT} = -\frac{1}{2T}\,h_n(T) - h_n(T) + h_{n+2}(T)$$

$$= -\frac{1}{2T}\,h_n(T) - ke^{-T} \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{n+1}{2} + i\right)T^{i-1/2}}{(n+2+2i)\Gamma\left(\frac{n+2}{2} + i\right)i!}.$$

It is apparent that for any positive n and any positive T, t or ρ^2, this derivative is negative. Consequently, $(1/t)\ E(\cos\theta)$ is a decreasing function of t, so that t should be as small as possible. The minimum number of trials required to estimate a gradient by means of uncorrelated estimates of its components is one more than the number of factors. As a result, $t = n + 1$ is the "optimum" number of trials to be used in each set of a maximum seeking experiment when such estimation is to be done.

The simplex designs are very appropriate for this kind of experimentation. These and other useful first order designs are discussed in the reference by Box. In those cases where the factors can be assigned only discrete levels, the designs given by Plackett and Burman can be used.

IV. *Efficiency of Replication*

It is of interest to see how efficient more than $n + 1$ trials in a set would be since other considerations may make it desirable to do as many trials as would be worthwhile within a single set. This would be the case, for example, if in some agricultural applications, many trials may be carried out simultaneously, but it takes an entire growing season to do any set of trials.

As a notational convenience, the number of trials for which $t = n + 1$ is designated as a single replicate, $R = 1$, and $t = R(n + 1)$ is an Rth replicate. Let G be the value of ρ when $R = 1$. Thus, G is the magnitude of the gradient relative to the standard error of estimation of a component when the number of trials used to do this estimation is $n + 1$. For those experimental designs for which the standard error of component estimation is reduced in proportion to $R^{-1/2}$, e.g. "pure" replication, there results the relation $\rho = R^{1/2}G$. The efficiency of

TABLE 2

Efficiency of R, the Number of Replicates, for n, the Number of Factors and G, the Magnitude of the Gradient Relative to the Standard Error of Estimation When $R = 1$

The Number of Factors is $n = 2$

	Relative Magnitude of the Gradient					
Replicates, R	0.25	0.5	1	2	4	∞
1	1.000	1.000	1.000	1.000	1.000	1.000
2	0.702	0.686	0.621	0.550	0.510	0.500
4	0.489	0.459	0.369	0.286	0.257	0.250
8	0.334	0.292	0.203	0.146	0.129	0.125
16	0.224	0.174	0.106	0.073	0.065	0.0625

The Number of Factors is $n = 8$

	Relative Magnitude of the Gradient, G					
Replicates, R	0.25	0.5	1	2	4	∞
1	1.000	1.000	1.000	1.000	1.000	1.000
2	0.705	0.699	0.676	0.618	0.547	0.500
4	0.495	0.482	0.411	0.358	0.287	0.250
8	0.346	0.326	0.272	0.196	0.149	0.125
16	0.239	0.213	0.158	0.103	0.076	0.0625

The Number of Factors is $n = 32$

	Relative Magnitude of the Gradient, G					
Replicates, R	0.25	0.5	1	2	4	∞
1	1.000	1.000	1.000	1.000	1.000	1.000
2	0.706	0.704	0.697	0.672	0.614	0.500
4	0.499	0.495	0.480	0.435	0.354	0.250
8	0.351	0.345	0.322	0.267	0.196	0.125
16	0.247	0.237	0.209	0.154	0.105	0.0625

an Rth replicate for particular n and G is defined as the improvement per unit of effort expected as a result of the Rth replicate design as compared with the improvement per unit of effort expected when $R = 1$:

$$\text{Efficiency of} \quad (R \mid n, G) = \frac{\frac{1}{R} E(\cos\theta \mid n, \rho = \sqrt{R}G)}{E(\cos\theta \mid n, \rho = G)}.$$

REFERENCES

1. Box, G. E. P. [1952]. Multifactor designs of first order. *Biometrika 39*, 49–57.
2. Box, G. E. P., and Wilson, K. B. [1951]. On the experimental attainment of optimum conditions. *Jour. Roy. Stat. Soc., 13*, 1–45.
3. Brooks, S. H. [1959]. A comparison of maximum seeking methods. *Operations Research 7*, 430–57.
4. Plackett, R. L., and Burman, J. P. [1946]. The Design of optimum multifactorial experiments. *Biometrika 33*, 305–25.

A STOCHASTIC STUDY OF THE LIFE TABLE AND ITS APPLICATIONS

III. THE FOLLOW-UP STUDY WITH THE CONSIDERATION OF COMPETING RISKS[1,2]

CHIN LONG CHIANG
University of California, Berkeley, California, U.S.A.

INTRODUCTION

Statistical studies falling into the general category of life testing and medical follow-up have as their common immediate objective the estimation of life expectation and survival rates for a defined population at risk. Usually such a study must be brought to a close before all the information on survival (of patients, electric bulbs, automobiles, etc.) is complete, and thus the study is said to be truncated. Whether the investigation is basically concerned with life testing or with medical follow-up, the nature of the problem is the same, although differences in sample size may call for different approaches. Thus methods developed for life testing may be applied to follow-up studies when the underlying conditions are met, and vice versa. In this study cancer survival data utilizing a large sample will be used as illustrative material, and we shall accordingly use the terminology of the medical follow-up study as a matter of convenience.

We are concerned then with a typical follow-up study in which a group of individuals with some common morbidity experience are followed from a well-defined zero point, such as date of hospital admission. Perhaps we wish to evaluate a certain therapeutic measure by comparing the expectation of life and survival rates of treated and untreated patients. Or we may wish to compare the expectation of life of treated and presumably cured patients with that of normal persons. When the period of observation is ended, there will usually remain a number of individuals on whom the mortality data in a typical study will be incomplete. Of first importance among these are the

[1] This study was completed while the author was a Special Research Fellow of the National Heart Institute, Public Health Service, U. S. Department of Health, Education and Welfare.

[2] Parts of this paper were presented at the joint meeting of the American Statistical Association, the Institute of Mathematical Statistics, and the Biometric Society in Washington, D. C., December 29, 1959.

persons still alive at the close of the study. Secondly, if the investigation is concerned with mortality from a specific cause, the necessary information is incomplete and unavailable for patients who died from other causes. In addition, there will usually be a third group of patients who were "lost" to the study because of follow-up failure. These three groups present a number of statistical problems in the estimation of the expectation of life and survival rates. Many significant studies have been made along these lines, among them the early works of Greenwood [13] and Karn [17], the actuarial method of Berkson and Gage [1], a stochastic model of competing risks by Fix and Neyman [12], the parametric studies by Berkson and Gage [2] and Boag [3], the non-parametric approach of Kaplan and Meier [16], studies on life testing by Epstein and Soble [10] and other interesting works by Dorn [8], Elveback [9], Fix [11], Harris, Meier and Tukey [14], and Littell [18].

The purpose of this paper is to adapt the biometric functions of the life table to the special conditions of the follow-up study. Part I considers the general type of study in which survival experience is investigated without specification as to the cause of death. An exact formula will be presented for the maximum likelihood estimator of the probability of death and its asymptotic variance. Special attention will be given to a method for computing the observed expectation of life in truncated studies and the corresponding variance. In Part II the discussion will be extended to apply to studies of mortality from a specific cause in the presence of competing risks. The relations between net, crude, and partial crude probabilities will be reviewed and formulas developed for their estimators and the corresponding variances and covariances. In the last part of the paper, data obtained from the California State Department of Public Health will be used to illustrate the application of the theoretical matter presented in Parts I and II.

Throughout this paper we shall assume that all individuals in a sample are subject to the same force of mortality (or, instantaneous death probability), and that the probability of dying for one individual is not influenced by the death of any other individual in the group. This is to say that the life-times of all individuals in a group are treated as independent and identically distributed random variables. We shall also assume in Part I that there will be no individuals lost to observation because of follow-up failure. The problem of lost cases will be considered in Part II (Remark 3).

For simplicity of presentation, a constant time interval (year) will be used. However, the methods developed in this paper apply equally well to cases where intervals are of different lengths; although the observed expectation of life will have a slightly different form (Cf. [5] and [6]).

The probability symbols used in this paper are listed below for convenient reference. Considering death without specification to cause:

p_x = Pr [an individual alive at time x will survive the interval $(x, x + 1)$],

q_x = Pr [an individual alive at time x will die in the interval $(x, x + 1)$],

and obviously $p_x + q_x = 1$. When death is studied by cause, or risk, we have the net probabilities:

q_{xk} = Pr [an individual alive at time $x \cdot$ will die in the interval $(x, x + 1)$ if risk R_k is the only acting risk of death in the population],

$q_{x.k}$ = Pr [an individual alive at time x will die in the interval $(x, x + 1)$ if risk R_k is eliminated from the population];

the crude probability:

Q_{xk} = Pr [an individual alive at time x will die from cause R_k in the interval $(x, x + 1)$, in the presence of all other risks in the population];

and the partial crude probabilities:

$Q_{xk.1}$ = Pr [an individual alive at time x will die from cause R_k in the interval $(x, x + 1)$, when only risk R_1 is eliminated from the population],

$Q_{xk.12}$ = Pr [an individual alive at time x will die from cause R_k in the interval $(x, x + 1)$, when risks R_1 and R_2 are eliminated from the population].

PART I. THE ESTIMATION OF THE PROBABILITY OF SURVIVAL AND EXPECTATION OF LIFE

1.1. *The basic random variables and their joint probability function.*

Consider a follow-up study conducted over a period of y years. A total of N_0 individuals are accepted into the study at any time prior to the closing date[3] and are observed until death or until the study is terminated, whichever comes first. If we set the time of entrance into

[3] Although we have used the common closing date method to illustrate the techniques developed in this paper, it should be pointed out that these techniques are equally applicable to the date of last reporting method.

the study as the common point of origin for all N_0 individuals, then N_0 is taken to be the number with which the study began, or the number of individuals alive at time zero. Let x be the exact number of years since entrance into the study, and N_x the number of individuals who survive to the common point x. Clearly, N_x may also be defined as the number of survivors who entered the study at least x years before its closing date. The number of survivors will decrease as x increases, not only because of deaths but also because of withdrawals due to the closing of the study. We will describe this process of depletion systematically for the typical interval $(x, x + 1)$ with reference to Table 1.[4]

At time x, the N_x survivors who begin the interval can be divided into two mutually exclusive groups according to their date of entrance into the study. A group of m patients entered the study more than $x + 1$ years before the closing date of the study. Out of these, δ patients will die in the interval and s will survive to begin the next interval. The second group of n patients entered the study less than $x + 1$ years before its termination, and hence are all counted as withdrawals

TABLE 1

Distribution of N_x Patients According to Withdrawal Status, Survival Status, and Cause of Death in the Interval $(x, x + 1)$*

	Withdrawal status in the interval		
Survival status and cause of death	Total number of patients	Number not due for withdrawal**	Number due for withdrawal***
Total	N_x	m	n
Survivors	$s + w$	s	w
Deaths, all causes	D	δ	ϵ
Cause of death			
R_1	D_1	δ_1	ϵ_1
R_2	D_2	δ_2	ϵ_2
.	.	.	.
.	.	.	.
.	.	.	.
R_c	D_c	δ_c	ϵ_c

*The subscript x, which should be added to each of the symbols in the table, is deleted to simplify formulas in the text.

**Patients admitted to the study more than $(x + 1)$ years before closing date.

***Patients admitted to the study less than $(x + 1)$ years but more than x years before closing date.

[4] The second part of Table 1, death by cause, was included for use in Part II of this paper.

in the interval $(x, x + 1)$, whether or not they survive, because for them the closing date precedes their $(x + 1)$-anniversary date. Let us say that ϵ will die before the closing date and w will survive to withdraw alive.

Thus s, δ, w, and ϵ are the basic random variables (upper part of Table 1) whose distribution depends on the force of mortality. The values that these random variables take on will be used to estimate the probability p_x that a patient will survive the interval $(x, x + 1)$, and its complement q_x, the probability of death in the interval. The first step is to derive the joint probability function of these random variables.

Let μ_τ, a function of time τ, be the force of mortality acting on each individual in the study, such that

$$\mu_\tau \, \Delta\tau + o(\Delta\tau) = \text{Pr [an individual alive at time } \tau \text{ will die in the interval } (\tau, \tau + \Delta\tau)], \text{ for } \tau \geq 0,$$

where $\Delta\tau$ stands for an infinitesimal time interval and $o(\Delta\tau)$ a quantity of smaller order of magnitude than $\Delta\tau$. It can be shown [5] that

$$p_x(t) = \exp\left\{-\int_x^{x+t} \mu_\tau \, d_\tau\right\}$$

is the probability that an individual alive at x will survive to $(x + t)$. If we assume a constant force of mortality within the interval $(x, x + 1)$, say $\mu_\tau = \mu_x$ depending only on x, for $x < \tau \leqq x + 1$, then the probability of surviving the interval is given by[5]

$$p_x = e^{-\mu_x}.$$

For the subinterval $(x, x + t)$, with t between zero and one, we have

$$p_x(t) = e^{-t\mu_x} = p_x^t, \quad \text{for} \quad 0 < t \leq 1.$$

Consider first the group of m individuals, each of whom has a constant probability p_x of surviving and a probability $q_x = 1 - p_x$ of dying in the interval $(x, x + 1)$. We have then a typical binomial case with the probability function:

$$f_1 = p_x^s(1 - p_x)^\delta. \tag{1}$$

The expected number of survivors and deaths are given, respectively, by

$$E(s \mid m) = mp_x, \quad \text{and} \quad E(\delta \mid m) = m(1 - p_x). \tag{2}$$

The distribution of the random variables in the group due for withdrawal is not so straightforward. Making the assumption that, on

[5] When the assumption of a constant force of mortality is strong for an interval where the death rate is high, one may subdivide the interval and estimate the probability for each subinterval separately.

the average, each of the m individuals will withdraw at the point $x + \frac{1}{2}$, the probability of withdrawing alive is equal to $p_x^{1/2}$, and the probability of dying before the time of withdrawal $(1 - p_x^{1/2})$. Again we have a binomial case with the probability function[6]

$$f_2 = p_x^{w/2}(1 - p_x^{1/2})^{\epsilon}. \tag{3}$$

The expected number of survivors and deaths are given, respectively, by

$$E(w \mid n) = np_x^{1/2}, \quad \text{and} \quad E(\epsilon \mid n) = n(1 - p_x^{1/2}). \tag{4}$$

Since the N_x individuals are divided at time x into two distinctly different groups according to their withdrawal status, the joint probability of all the random variables is the product of the two probability functions (1) and (3),

$$f_1 f_2 = p_x^{(s+w/2)}(1 - p_x)^{\delta}(1 - p_x^{1/2})^{\epsilon}. \tag{5}$$

1.2. *The maximum likelihood estimators and their asymptotic variances.*

We are now in a position to use the maximum likelihood principle to obtain the estimator of the probability p_x and its complement q_x, and the asymptotic variance. Taking the logarithm of the joint probability function (5), we have the likelihood function

$$L = (s + \tfrac{1}{2}w) \ln p_x + \delta \ln (1 - p_x) + \epsilon \ln (1 - p_x^{1/2}). \tag{6}$$

Differentiating (6) with respect to p_x and setting the derivative equal to zero give the likelihood equation

$$(s + \tfrac{1}{2}w)\hat{p}_x^{-1} - \delta(1 - \hat{p}_x)^{-1} - \frac{\epsilon}{2}\,\hat{p}_x^{-1/2}(1 - \hat{p}_x^{1/2})^{-1} = 0,$$

which implies

$$\left(N_x - \frac{n}{2}\right)\hat{p}_x + \frac{\epsilon}{2}\,\hat{p}_x^{1/2} - (s + \tfrac{1}{2}w) = 0, \tag{7}$$

a quadratic equation in $\hat{p}_x^{1/2}$. Since $\hat{p}_x^{1/2}$ cannot take on negative values, we have the estimators,

$$\hat{p}_x = \left[\frac{-\frac{1}{2}\epsilon + \sqrt{\frac{1}{4}\epsilon^2 + 4(N_x - \frac{1}{2}n)(s + \frac{1}{2}w)}}{2(N_x - \frac{1}{2}n)}\right]^2 \tag{8}$$

[6] A more plausible assumption perhaps is that a withdrawal takes place randomly throughout the interval. However, under this assumption the probability of withdrawing alive is

$$\int_x^{x+1} e^{-(\tau - x)\mu_x}\, d\tau = -(1 - p_x)/\ln p_x,$$

and the resulting maximum likelihood equation is too unwieldy. When the probability of survival is not too low, the above expression and $p_x^{\frac{1}{2}}$ are close to each other. For $p_x = .70$, for example, $p_x^{\frac{1}{2}} = .837$ and $-(1 - p_x)/\ln p_x = .840$. In the case of extremely high mortality, one should subdivide the interval.

and

$$\hat{q}_x = 1 - \hat{p}_x \, . \tag{9}$$

The maximum likelihood estimator (8) is not unbiased; however, it is consistent in the sense that when the random variables s, w, and ϵ are replaced with their respective expectations as given by (2) and (4), the resulting expression is identical with the probability p_x. That is,

$$p_x \equiv \left[\frac{-\frac{1}{2}n(1 - p_x^{\frac{1}{2}}) + \sqrt{\frac{1}{4}n^2(1 - p_x^{\frac{1}{2}})^2 + 4(N_x - \frac{1}{2}n)(mp_x + \frac{1}{2}np_x^{\frac{1}{2}})}}{2(N_x - \frac{1}{2}n)}\right]^2 .$$

To derive the formula for the asymptotic variance of the estimator $\hat{p}_x$ (or $\hat{q}_x$), we find the expectation of the second derivative of (6):

$$E\left(\frac{\partial^2 L}{\partial p_x^2}\right) = -\left[\frac{M_x}{p_x q_x} + \pi\right], \tag{10}$$

where

$$M_x = m + n(1 + p_x^{1/2})^{-1} \tag{11}$$

and

$$\pi = \frac{n}{4(1 + p_x^{1/2})^2 p_x^{3/2}} (1 - p_x^{1/2}). \tag{12}$$

According to the theorem on the asymptotic efficiency of an estimator, the asymptotic variance of $\hat{p}_x$ (or $\hat{q}_x$) is given by the negative inverse of the expectation (10),

$$\sigma^2_{\hat{p}_x} = 1/[M_x/(p_x q_x) + \pi]. \tag{13}$$

Usually the quantity π in the denominator of (13) will be small in comparison with the preceding term, and may be neglected to give the approximate formula

$$\sigma^2_{\hat{p}_x} = p_x q_x / M_x \, . \tag{14}$$

The sample variance of $\hat{p}_x$ (or $\hat{q}_x$) is obtained by substituting (8) and (9) in (13) or (14).

Remark 1: The problem discussed here relates to the study in which N_x is large. If N_x is small, one may use the exact time of death of each of the D patients and the exact time of withdrawal of each of the w patients to estimate the probability p_x. In this case, there will be N_x individual observations within the interval $(x, x + 1)$, and obviously it is unnecessary to consider the N_x patients as two distinct groups according to their withdrawal status. Let $t_i \leq 1$ be the time of death within the interval $(x, x + 1)$ of the i-th death, for $i = 1, \cdots . D$, with a probability

$$e^{-t_i\mu_x}\mu_x\, dt_i\ , \quad \text{for} \quad i = 1, \cdots, D;$$

let $T_j \leq 1$ be the time of withdrawal alive of the j-th withdrawal, for $j = 1, \cdots , w$, with a probability

$$e^{-T_j\mu_x}, \quad \text{for} \quad j = 1, \cdots, w;$$

then the joint probability function of all the N_x observations becomes

$$p_x^s \prod_{i=1}^{D} (e^{-t_i\mu_x}\mu_x\, dt_i) \prod_{j=1}^{w} (e^{-T_j\mu_x}) = (-\ln p_x)^D p_x^{(s+\sum_{i=1}^{D} t_i+\sum_{j=1}^{w} T_j)} \prod_{i=1}^{D} dt_i\ .$$

Maximizing the last expression with respect to p_x gives the maximum-likelihood estimator (cf. [18]),

$$\hat{p}_x = \exp\left[-D\Big/\left(s + \sum_{i=1}^{D} t_i + \sum_{j=1}^{w} T_j\right)\right].$$

1.3. *Observed expectation of life.*

A life table for the follow-up subjects can be readily constructed, once $\hat{p}_x$ and $\hat{q}_x$ have been determined from (8) and (9) for each interval of the study period. Let an arbitrary number l_0 denote the number of patients admitted to the study. The number l_x who survive to the exact time x is computed from the formula $l_x = l_0\hat{p}_0\hat{p}_1 \cdots \hat{p}_{x-1}$, and $l_x/l_0 = \hat{p}_0\hat{p}_1 \cdots \hat{p}_{x-1}$ is the estimated x-year survival rate. For a patient alive at time x, the observed expectation of life can be expressed by the equation:

$$\hat{e}_x = a_x + c_{x+1}\hat{p}_x + c_{x+2}\hat{p}_x\hat{p}_{x+1} + \cdots \\ + c_y\hat{p}_x\hat{p}_{x+1} \cdots \hat{p}_{y-1} + c_{y+1}\hat{p}_x\hat{p}_{x+1} \cdots \hat{p}_y + \cdots , \qquad (15)$$

where a_x is the average time lived in the interval $(x, x + 1)$ by the patients who die in that interval, and $c_x = 1 - a_{x-1} + a_x$. If, in a study covering a period of y years, there are no survivors remaining from the patients who entered the study in its first year, $\hat{p}_{y-1}$ will be zero, and $\hat{e}_x$ can be computed readily from the collected data. In the typical study, however, there will be w_{y-1} survivors who entered the study in its first year and withdraw alive in the final interval $(y - 1, y)$. In such cases, it is evident from (8) that $\hat{p}_{y-1}$ is greater than zero and the values of $\hat{p}_y$, $\hat{p}_{y+1}$, $\cdots$ are not observed within the limits of the study. Consequently, $\hat{e}_x$ cannot be computed from equation (15).

It is nevertheless possible to estimate $\hat{e}_x$ with a certain degree of accuracy if w_{y-1} is small. Suppose we rewrite equation (15) in the form

$$\hat{e}_x = a_x + c_{x+1}\hat{p}_x + c_{x+2}\hat{p}_x\hat{p}_{x+1} + \cdots + c_y\hat{p}_x\hat{p}_{x+1} \cdots \hat{p}_{y-1} + \frac{l_y}{l_x}(c_{y+1}\hat{p}_y + c_{y+2}\hat{p}_y\hat{p}_{y+1} + \cdots), \qquad (16)$$

where l_y/l_x is written for $\hat{p}_x\hat{p}_{x+1} \cdots \hat{p}_{y-1}$. The problem is to estimate the values of $\hat{p}_y$, $\hat{p}_{y+1}$, $\cdots$ in the last term, since the preceding terms can be computed from the data available.

As a first approach, consider a typical interval $(z, z + 1)$ beyond time y with the probability of surviving the interval:

$$p_z = \exp\left(-\int_z^{z+1} \mu_\tau \, d\tau\right), \quad \text{for} \quad z = y, y + 1, \cdots .$$

If the force of mortality is constant for $z \geq y$, the probability of survival is independent of z and we may write

$$p_z = e^{-\mu} = p, \quad \text{for} \quad z = y, y + 1, \cdots .$$

Under this assumption $c_z = 1$, and we may replace the last term of (16) with $(l_y/l_x)(\hat{p} + \hat{p}^2 + \cdots)$, which converges to $(l_y/l_x)\hat{p}/(l - \hat{p})$. As a result, we have

$$\hat{e}_x = a_x + c_{x+1}\hat{p}_x + c_{x+2}\hat{p}_x\hat{p}_{x+1} + \cdots + c_y\hat{p}_x\hat{p}_{x+1} \cdots \hat{p}_{y-1} + \frac{l_y}{l_x}[\hat{p}/(1 - \hat{p})].$$

Clearly, $\hat{p}$ may be set equal to $\hat{p}_{y-1}$ if the force of mortality is assumed to be constant beginning with time $(y - 1)$ instead of time y. From the point of view of sample variation, however, it is desirable to base the estimate $\hat{p}$ on as large a sample size as possible. Suppose there exists a time T, for $T < y$, such that $\hat{p}_T$, $\hat{p}_{T+1}$, $\cdots$ are approximately equal, thus indicating a constant force of mortality after time T. Then $\hat{p}$ may be set equal to $\hat{p}_T$, and the formula for the observed expectation of life becomes (cf. [4]):

$$\hat{e}_x = a_x + c_{x+1}\hat{p}_x + c_{x+2}\hat{p}_x\hat{p}_{x+1} + \cdots + c_y\hat{p}_x\hat{p}_{x+1} \cdots \hat{p}_{y-1} + \frac{l_y}{l_x}[\hat{p}_T/(1 - \hat{p}_T)], \qquad (17)$$

for $x = 0, \cdots, y - 1$. When a_x is approximated with $\frac{1}{2}$, $c_x = 1$.

Although formula (17) holds for $x = 0, \cdots, y - 1$, it will be apparent to the reader that the smaller the x, the larger the value of l_x, and the smaller will be the contribution of the last term. If the ratio $1_y/l_x$ is small, the error in assuming a constant force of mortality beyond y and in the choice of $\hat{p}_T$ will have but little effect on the value of $\hat{e}_x$.

1.4. *The sample variance of the observed expectation of life.*

To avoid confusion in notation let us denote by α a fixed number and consider the observed expectation of life $\hat{e}_\alpha$ as given in formula (17). It was proven in [5] that the estimated probabilities of surviving any two non-overlapping intervals have a zero covariance, and hence the sample variance of the observed expectation of life may be computed from

$$S^2_{\hat{e}_\alpha} = \sum_{x \geq \alpha} [\partial \hat{e}_\alpha / \partial \hat{p}_x]^2 S^2_{\hat{q}_x} , \tag{18}$$

where the derivatives are taken at the observed point $\hat{p}_x$, for $x \geq \alpha$. In the present case, we have

$$[\partial \hat{e}_\alpha / \partial \hat{p}_x] = \hat{p}_{\alpha x}[\hat{e}_{x+1} + (1 - a_x)], \quad \text{for} \quad x \neq T, \tag{19}$$

and

$$\begin{aligned}[\partial \hat{e}_\alpha / \partial \hat{p}_T] &= \hat{p}_{\alpha T}[\hat{e}_{T+1} + (1 - a_T) + \{\hat{p}_{TY}/(1 - \hat{p}_T)^2\}], \text{ for } \alpha \leq T, \\ &= \hat{p}_{\alpha y}/(1 - \hat{p}_T)^2, \quad \text{for} \quad \alpha > T,\end{aligned} \tag{20}$$

where $\hat{p}_{\alpha x} = \hat{p}_\alpha \hat{p}_{\alpha+1} \cdots \hat{p}_{x-1}$. Substituting (19) and (20) in (18) gives the sample variance of $\hat{e}_\alpha$,

$$\begin{aligned}S^2_{\hat{e}_\alpha} = \sum_{\substack{x=\alpha \\ x \neq T}}^{y-1} \hat{p}^2_{\alpha x}[\hat{e}_{x+1} + (1 - a_x)]^2 S^2_{\hat{q}_x} \\ + \hat{p}^2_{\alpha T}[\hat{e}_{T+1} + (1 - a_T) + \{\hat{p}_{TY}/(1 - \hat{p}_T)^2\}]^2 S^2_{\hat{q}_T}\end{aligned} \tag{21}$$

for $\alpha \leq T$, and

$$\begin{aligned}S^2_{\hat{e}_\alpha} = \sum_{x=\alpha}^{y-1} \hat{p}^2_{\alpha x}[\hat{e}_{x+1} + (1 - a_x)]^2 S^2_{\hat{q}_x} \\ + \{\hat{p}^2_{\alpha y}/(1 - \hat{p}_T)^4\} S^2_{\hat{q}_T} , \quad \text{for} \quad \alpha > T.\end{aligned} \tag{21a}$$

The value of $\hat{p}_x$ and the sample variance of $\hat{q}_x$ are obtained from formulas (8) and (13), respectively. When a_x is approximated with $\frac{1}{2}$, the quantity $(1 - a_x)$ in formulas (21) and (21a) may be replaced by $\frac{1}{2}$.

PART II. CONSIDERATION OF COMPETING RISKS

2.1. *Relations between net, crude, and partial crude probabilities.*

In a follow-up study, as in general mortality analysis, one may be interested in death due to a specific cause, or to a group of causes.

Depending upon the questions to be answered, the investigator may explore three general types of probabilities of death with respect to a specific cause, or risk:

1. *The crude probability.* The probability of death from a specific cause in the presence of all other risks in a population.
2. *The net probability.* The probability of death if a specific cause were the only cause in effect in the population or, conversely, the probability of death if a specific risk were eliminated from the population.
3. *The partial crude probability.* The probability of death from a specific cause in the presence of all other risks but with a second risk eliminated from the population.

Obviously, in the human population, the net and partial crude probabilities usually cannot be estimated directly except through their relations with the crude probability. The study of such relations is part of the problem of "competing risks", or "multiple-decrement". The subject has been variously discussed in the literature (see, for example, [12], [15], and [19]) and will be reviewed here only by way of introducing notation.

Assume c risks of death (or causes) acting simultaneously on each individual of a population (that is, competing for the life of the individual), and let these risks be denoted by $R_1, \cdots, R_c$. For each risk there is a corresponding force of mortality, $\nu_{\tau 1}, \cdots, \nu_{\tau c}$, each of which is a function of time τ, and the sum of these

$$\nu_{\tau 1} + \cdots + \nu_{\tau c} = \mu_\tau \tag{22}$$

is then the total force of mortality. Within the time interval $(x, x + 1)$, we shall assume a constant force of mortality for each risk, say $\nu_{\tau k} = \nu_{xk}$, depending only on x and k, for $x < \tau \leq x + 1$. For all risks, we have $\mu_\tau = \mu_x$, for $x < \tau \leq x + 1$.

Let $Q_{xk}(t)$ be the crude probability that an individual alive at time x will die in the interval $(x, x + t)$, for $0 < t \leq 1$, from cause R_k in the presence of all other risks. It follows directly from addition and multiplication theorems that

$$Q_{xk}(t) = \int_x^{x+t} e^{-(\tau - x)\mu_x} \nu_{xk}\, d\tau, \quad \text{for} \quad 0 < t \leq 1; \qquad k = 1, \cdots, c. \tag{23}$$

The first factor of the integrand is the probability of surviving from x to τ when all risks of death are acting, while the second factor is the instantaneous probability of death from cause R_k. Integrating (23) gives

$$Q_{xk}(t) = \frac{\nu_{xk}}{\mu_x}[1 - e^{-t\mu_x}] = \frac{\nu_{xk}}{\mu_x}[1 - p_x(t)],$$
$$\text{for} \quad 0 < t \leq 1; \qquad k = 1, \cdots, c. \tag{24}$$

It is clear from (22) that the sum of the crude probabilities in (24) is equal to the complement of $p_x(t)$, or

$$Q_{x1}(t) + \cdots + Q_{xc}(t) + p_x(t) = 1, \quad \text{for} \quad 0 < t \leq 1. \tag{25}$$

For $t = 1$, we shall abbreviate $Q_{xk}(1)$ to Q_{xk}, etc. For the purpose of this study we are particularly interested in the subinterval $(x, x + \frac{1}{2})$, with

$$Q_{xk}(\tfrac{1}{2}) = \frac{\nu_{xk}}{\mu_x}[1 - e^{-\mu_x/2}] = Q_{xk}[1 + p_x^{1/2}]^{-1}, \quad \text{for} \quad k = 1, \cdots, c. \tag{26}$$

In this case the sum of the crude probabilities (26) is the complement of $p_x^{1/2}$, the probability of surviving half the interval, and

$$Q_{x1}[1 + p_x^{1/2}]^{-1} + \cdots + Q_{xc}[1 + p_x^{1/2}]^{-1} + p_x^{1/2} = 1. \tag{27}$$

When risk R_k acts alone, the net probability that an individual alive at time x will die in the interval $(x, x + 1)$ is

$$q_{xk} = 1 - e^{-\nu_{xk}} = 1 - [e^{-\mu_x}]^{\nu_{xk}/\mu_x}. \tag{28}$$

From formulas (24) and (28) we obtain the relation between the net and crude probabilities

$$q_{xk} = 1 - p_x^{Q_{xk}/q_x}, \quad \text{for} \quad k = 1, \cdots, c. \tag{29}$$

By analogy we can write the net probability of death in the interval $(x, x + 1)$ when risk R_k is eliminated,

$$q_{x.k} = 1 - p_x^{(q_x - Q_{xk})/q_x}, \quad \text{for} \quad k = 1, \cdots, c. \tag{30}$$

Now suppose that R_1 is eliminated as a cause of death, and let $Q_{xk.1}$ be the partial crude probability that an individual alive at time x will die in the interval $(x, x + 1)$ from cause R_k in the presence of all other risks, for $k = 2, \cdots, c$. Using a similar reasoning as in the crude probability [eq. (23)], we can write

$$Q_{xk.1} = \int_x^{x+1} e^{-(\tau - x)(\mu_x - \nu_{x1})} \nu_{xk} \, d\tau$$
$$= \frac{Q_{xk}}{q_x - Q_{x1}}[1 - p_x^{(q_x - Q_{x1})/q_x}], \quad \text{for} \quad k = 2, \cdots, c. \tag{31}$$

Similarly, if risks R_1 and R_2 are eliminated, the partial crude probability that an individual alive at x will die in the interval $(x, x + 1)$ from cause R_k in the presence of all other causes is given by

$$Q_{xk \cdot 12} = \frac{Q_{xk}}{q_x - Q_{x1} - Q_{x2}} [1 - p_x^{(q_x - Q_{x1} - Q_{x2})/q_x}], \quad \text{for } k = 3, \cdots, c. \tag{32}$$

A detailed discussion on the partial crude probabilities is given in [7].

2.2. *The basic random variables and their joint probability function.*

The identification of the random variables in a follow-up study in the presence of competing risks and the derivation of their joint probability function follows directly from the discussion in Section 1 of Part I. The deaths in each of the two groups according to withdrawal status are further divided by cause of death as shown in Table 1, Part I.

Each of the m individuals not due for withdrawal in the interval $(x, x + 1)$ will fall into one of the $c + 1$ mutually exclusive groups, depending upon whether he survives the interval or dies from cause R_1, $\cdots$, R_c, with the sum of the corresponding probabilities p_x, Q_{x1}, $\cdots$, Q_{xc}, equal to unity [eq. (25)]. Thus we have a multinomial case with the probability function

$$f_1 = p_x^s Q_{x1}^{\delta_1} \cdots Q_{xc}^{\delta_c}, \tag{33}$$

where s is the number of survivors and δ_k is the number of deaths from cause R_k, for $k = 1, \cdots, c$. Their mathematical expectations are given respectively, by

$$E(s \mid m) = mp_x, \quad \text{and} \quad E(\delta_k \mid m) = mQ_{xk}, \quad \text{for} \quad k = 1, \cdots, c. \tag{34}$$

Each individual in the group of n due for withdrawal in the interval $(x, x + 1)$ has the probability $p_x^{1/2}$ of withdrawing alive and the probability

$$Q_{xk}(\tfrac{1}{2}) = Q_{xk}(1 + p_x^{1/2})^{-1}, \quad \text{for} \quad k = 1, \cdots, c, \tag{26}$$

of dying from cause R_k. Since $p_x^{1/2}$ and the probabilities in (26) again add up to unity [eq. (27)], the n observations also constitute a multinomial case with the probability function (cf. footnote 6)

$$f_2 = p_x^{w/2} \prod_{k=1}^{c} [Q_{xk}(1 + p_x^{1/2})^{-1}]^{\epsilon_k}, \tag{35}$$

where w is the number of individuals withdrawing alive and ϵ_k the number of individuals who die from cause R_k before the time of withdrawal. The mathematical expectations are, respectively,

$$E(w \mid n) = np_x^{1/2}, \quad \text{and} \quad E(\epsilon_k \mid n) = nQ_{xk}(1 + p_x^{1/2})^{-1}, \quad \text{for} \quad k = 1, \cdots, c. \tag{36}$$

Because of the separation of the individuals into two distinct groups at time x according to their withdrawal status, the joint probability of all the random variables in Table 1 is the product of the two joint probabilities (33) and (35):

$$f_1 f_2 = p_x^{s+w/2} \prod_{k=1}^{c} Q_{xk}^{\delta_k} \prod_{k=1}^{c} [Q_{xk}(1 + p_x^{1/2})^{-1}]^{\epsilon_k}. \tag{37}$$

Formula (37) may be simplified by rearranging terms and using the relations, $D_k = \delta_k + \epsilon_k$, and $\epsilon = \epsilon_1 + \cdots + \epsilon_c$, to give the final form of the joint probability function

$$f_1 f_2 = p_x^{s+w/2}(1 + p_x^{1/2})^{-\epsilon} \prod_{k=1}^{c} Q_{xk}^{D_k} . \tag{38}$$

2.3. *Maximum-likelihood estimators of crude, net, and partial crude probabilities.*

We will again use the maximum likelihood principle to obtain the estimators of the probabilities: p_x, Q_{x1}, $\cdots$, Q_{xc}. In this case the likelihood function obtained from (38) is

$$L = (s + \tfrac{1}{2}w) \ln p_x - \epsilon \ln (1 + p_x^{1/2}) + D_1 \ln \left(1 - p_x - \sum_{k=2}^{c} Q_{xk}\right) + \sum_{k=2}^{c} D_k \ln Q_{xk} , \tag{39}$$

where the substitution

$$Q_{x1} = 1 - p_x - \sum_{k=2}^{c} Q_{xk} \tag{40}$$

has been made. Differentiating the likelihood function (39) with respect to p_x, Q_{x2}, $\cdots$, Q_{xc}, respectively, and setting the derivatives equal to zero, we obtain a system of c simultaneous equations:

$$\partial L/\partial p_x = [(s + \tfrac{1}{2}w)/\hat{p}_x] - \epsilon/[2\hat{p}_x^{1/2}(1 + \hat{p}_x^{1/2})] - (D_1/\hat{Q}_{x1}) = 0, \tag{41a}$$

$$\partial L/\partial Q_{xk} = (D_k/\hat{Q}_{xk}) - (D_1/\hat{Q}_{x1}) = 0, \quad \text{for} \quad k = 2, \cdots, c. \tag{41b}$$

From (41b) it can be deduced that

$$D_k/\hat{Q}_{xk} = D/(1 - \hat{p}_x), \quad \text{for} \quad k = 1, \cdots, c, \tag{42}$$

and therefore the ratio $D_1/\hat{Q}_{x1}$ in equation (41a) can be replaced with $D/(1 - \hat{p}_x)$. When this substitution is made and the terms in (41a) are rearranged we have a quadratic equation in $\hat{p}_x^{1/2}$ that is identical to equation (7) in Part I. Hence the estimators $\hat{p}_x$, in (8), and $\hat{q}_x$ for all causes of death will have the same value as in the simple case where death is investigated without specification to cause, as one would anticipate. Substituting (8) in (42) gives the estimators

$$\hat{Q}_{xk} = (D_k/D)\hat{q}_x\,, \quad \text{for} \quad k = 1, \cdots, c. \tag{43}$$

To obtain the estimators of the net and partial crude probabilities, we substitute (43) in formula (29), (30), (31), and (32), and after simplification,

$$\hat{q}_{xk} = 1 - \hat{p}_x^{D_k/D}, \quad \text{for} \quad k = 1, \cdots, c; \tag{44}$$

$$\hat{q}_{x.k} = 1 - \hat{p}_x^{(D-D_k)/D}, \quad \text{for} \quad k = 1, \cdots, c; \tag{45}$$

$$\hat{Q}_{xk.1} = [D_k/(D - D_1)][1 - \hat{p}_x^{(D-D_1)/D}], \quad \text{for} \quad k = 2, \cdots, c; \tag{46}$$

and

$$\hat{Q}_{xk.12} = [D_k/(D - D_1 - D_2)][1 - \hat{p}_x^{(D-D_1-D_2)/D}], \text{ for } k = 3, \cdots, c. \tag{47}$$

The estimators given in formulas (44), (45), (46), and (47) are also maximum likelihood estimators because of the invariance property of maximum likelihood estimators.

Remark 2: If there were no withdrawals in the interval $(x, x + 1)$, i.e., if $n = 0$, the problem is reduced to the classical multiple-decrement problem, with s survivors and $D_k = \delta_k$ deaths from cause R_k, for $k = 1, \cdots, c$. These random variables will still have a multinomial distribution [eq. (33)], and the formulas for the estimators of p_x, q_x, and Q_{xk} are reduced to

$$\hat{p}_x = s/N_x\,, \tag{8a}$$

$$\hat{q}_x = D/N_x\,, \tag{9a}$$

and

$$\hat{Q}_{xk} = D_k/N_x\,, \quad \text{for} \quad k = 1, \cdots, c. \tag{43a}$$

Formulas (44) through (47) may still be used for the estimators of the net and partial crude probabilities, but with $\hat{p}_x$ given by (8a).

Remark 3: The problem of cases lost to the study due to failure of follow-up is still unsolved, and perhaps it has no unique solution. Since the probability that a patient will be lost to follow-up is in part dependent upon the type of a study, assumptions with respect to lost cases may be valid for one study but not for another. If the number of lost cases is small. depending upon the type of study, one of the following assumptions may be made and the data handled accordingly: (1) patients lost will have the same probability of surviving as patients not lost, and may be deleted from the study; (2) all lost cases survive to the close of the study; (3) all die at the time of becoming lost; and (4) becoming lost is another competing risk. If sufficient knowledge of follow-up is unavailable, the fourth alternative is preferred.

2.4. *Asymptotic variance and covariance of the estimators.*

Formulas for the variance and covariance of the estimators may be determined by using the asymptotic property of maximum-likelihood estimators. The inverse of the asymptotic covariance matrix of the estimators, $\hat{p}_x$, $\hat{Q}_{x2}$, $\cdots$, $\hat{Q}_{xc}$, is given by

$$\|\Lambda\| = \left[\begin{array}{c|c} \left\| -E\left(\frac{\partial^2 L}{\partial p_x^2}\right)\right\| & \left\| -E\left(\frac{\partial^2 L}{\partial p_x\,\partial Q_{xk}}\right)\right\| \\ 1\times 1 & 1\times(c-1) \\ \hline \left\| -E\left(\frac{\partial^2 L}{\partial p_x\,\partial Q_{xk}}\right)\right\|' & \left\| -E\left(\frac{\partial^2 L}{\partial Q_{xh}\,\partial Q_{xk}}\right)\right\| \\ (c-1)\times 1 & (c-1)\times(c-1) \end{array}\right] \tag{48}$$

in which the elements are obtained by differentiating formula (41a) and (41b). Direct calculation gives the following mathematical expectations

$$-E(\partial^2 L/\partial p_x^2) = M_x[(1/p_x) + (1/Q_{x1})] + \pi, \tag{49}$$

$$-E(\partial^2 L/\partial p_x\,\partial Q_{xk}) = M_x/Q_{x1}\,, \quad \text{for} \quad k = 2, \cdots, c, \tag{50}$$

$$-E(\partial^2 L/\partial Q_{xh}\,\partial Q_{xk}) = M_x/Q_{x1}\,, \quad \text{for} \quad h \neq k; h, k = 2, \cdots, c, \tag{51}$$

and

$$-E(\partial^2 L/\partial Q_{xk}^2) = M_x[(1/Q_{x1}) + (1/Q_{xk})], \quad \text{for} \quad k = 2, \cdots, c, \tag{52}$$

where M_x and π are defined by equation (11) and (12), respectively, of Part I. Substituting the respective expectations into (48), we have

$$\|\Lambda\| = \left[\begin{array}{c|cccc} M_x\left(\frac{1}{p_x}+\frac{1}{Q_{x1}}\right)+\pi & M_x/Q_{x1} & M_x/Q_{x1} & \cdots & M_x/Q_{x1} \\ \hline M_x/Q_{x1} & M_x\left(\frac{1}{Q_{x1}}+\frac{1}{Q_{x2}}\right) & M_x/Q_{x1} & \cdots & M_x/Q_{x1} \\ M_x/Q_{x1} & M_x/Q_{x1} & M_x\left(\frac{1}{Q_{x1}}+\frac{1}{Q_{x3}}\right) & \cdots & M_x/Q_{x1} \\ \vdots & \vdots & \vdots & & \vdots \\ M_x/Q_{x1} & M_x/Q_{x1} & M_x/Q_{x1} & \cdots & M_x\left(\frac{1}{Q_{x1}}+\frac{1}{Q_{xc}}\right) \end{array}\right] \tag{53}$$

with its determinant

$$\Lambda = (M_x^c/Q_{x1} \cdots Q_{xc}p_x) + \pi(M_x^{c-1}/Q_{x1} \cdots Q_{xc})(1 - p_x). \qquad (54)$$

Denoting the cofactors of the determinant by Λ_{hk}, for $h, k = 1, \cdots, c$, the formulas for the asymptotic variance and covariance are given by

$$\sigma^2_{\hat{p}_x} = \Lambda_{11}/\Lambda = (p_xq_x/M_x)[1/\{1 + (\pi p_xq_x/M_x)\}], \qquad (55)$$

$$\sigma^2_{\hat{Q}_{xk}} = \Lambda_{kk}/\Lambda = [Q_{xk}(1 - Q_{xk})/M_x]\left[\frac{1 + \{\pi p_x(q_x - Q_{xk})/M_x(1 - Q_{xk})\}}{1 + \{\pi p_xq_x/M_x\}}\right],$$
$$\text{for} \quad k = 2, \cdots, c, \qquad (56)$$

$$\sigma^2_{\hat{p}_x, \hat{Q}_{xk}} = \Lambda_{1k}/\Lambda = -(p_xQ_{xk}/M_x)[1/\{1 + (\pi p_xq_x/M_x)\}],$$
$$\text{for} \quad k = 2, \cdots, c, \qquad (57)$$

and

$$\sigma^2_{\hat{Q}_{xh}, \hat{Q}_{xk}} = \Lambda_{xk}/\Lambda = -(Q_{xh}Q_{xk}/M_x)[\{1 + (\pi p_x/M_x)\}/\{1 + (\pi p_xq_x/M_x)\}],$$
$$\text{for} \quad h \neq k; \qquad h, k = 2, 3, \cdots, c. \qquad (58)$$

Since the term Q_{x1} was not explicitly included in the likelihood function (39), the formulas for the variance of $\hat{Q}_{x1}$ and the covariances between $\hat{Q}_{x1}$ and other estimators were not presented. It is obvious by reason of symmetry, however, that expressions for $\hat{Q}_{xk}$ do start from $\hat{Q}_{x1}$, which is to say that formulas (56), (57), and (58) hold also for $k = 1$.

The quantities inside the square brackets in formulas (55) through (58) may be approximated with unity when M_x is moderately large. These formulas then reduce to familiar expressions of the multinomial case:

$$\sigma^2_{\hat{p}_x} = p_xq_x/M_x, \qquad (59)$$

$$\sigma^2_{\hat{Q}_{xk}} = Q_{xk}(1 - Q_{xk})/M_x, \quad \text{for} \quad k = 1, \cdots, c, \qquad (60)$$

$$\sigma_{\hat{p}_x, \hat{Q}_{xk}} = -p_xQ_{xk}/M_x, \quad \text{for} \quad k = 1, \cdots, c, \qquad (61)$$

and

$$\sigma_{\hat{Q}_{xh}, \hat{Q}_{xk}} = -Q_{xh}Q_{xk}/M_x, \quad \text{for} \quad h \neq k; \qquad h, k = 1, 2, \cdots, c. \qquad (62)$$

Formulas for the asymptotic variance and covariance of the estimators of the net and partial crude probabilities can be obtained with the same approach as employed in [7]. To save space, only two formulas are presented below.

$$\sigma^2_{\hat{q}_{x \cdot k}} = [(1 - q_{x.k})^2/M_xp_xq_x]$$
$$\cdot [p_x \ln (1 - q_{xk}) \ln (1 - q_{x.k}) + (q_x - Q_{xk})^2], \qquad (63)$$

for $k = 1, \cdots, c$, for the net probability of death when risk R_k is eliminated, and

$$\begin{aligned}\sigma^2_{\hat{Q}xk.1} = {} & [(q_x - Q_{x1} - Q_{xk})/\{M_x(q_x - Q_{x1})Q_{xk}\}]Q^2_{xk.1} \\ & + [\{Q_{xk.1}(q_x - Q_{x1}) - Q_{xk}\}^2/\{M_x p_x q_x(q_x - Q_{x1})\}] \\ & \cdot[(q_x - Q_{x1}) + Q_{x1}p_x(\ln p_x/q_x)^2],\end{aligned} \tag{64}$$

for $k = 2, \cdots, c$, for the partial crude probability.

AN EXAMPLE OF LIFE TABLE CONSTRUCTION FOR THE FOLLOW-UP POPULATION

The application of the methods developed in Parts I and II will be illustrated with data collected by the Tumor Registry of the California State Department of Public Health. The material selected consists of 5982 patients[7] admitted to certain California hospitals and clinics between January 1, 1942, and December 31, 1954, with a diagnosis of cancer of the cervix uteri. For the purpose of this illustration, the latter date is taken as the common closing date of the study; the date of entrance to follow-up for each patient is the date of hospital admission.

The first step is to construct a table similar to Table 2, showing the survival experience of the patients grouped according to their withdrawal status. The interval length selected (column 1) will depend upon the nature of the investigation; in this case a fixed length of one year was convenient and satisfactory. The total number of patients admitted to the study is entered as N_0 in the first line of column 2, which in this example is 5982. To determine their withdrawal status in the first interval (0, 1) the patients were separated into two groups: admissions before 1954, and consequently at least one year before the close of the study; and admissions during the year 1954, all due for withdrawal since the study was terminated before their first anniversary. Of the patients admitted prior to 1954, s_0 (4030 in column 3) survived to their first anniversary and δ_0 (1287) died during the first year. The deaths were further divided by cause into δ_{01} (1105) deaths due to cancer of the cervix uteri and δ_{02} (182) deaths from all other causes. The survival status of the 1954 admissions is determined at the close of the study, as it is for patients due for withdrawal in any interval. In this study, w_0 (576) patients withdrew alive in the first interval, and ϵ_0 (89) patients died before the closing date. These deaths were again divided by cause into ϵ_{01} (70, column 9) and ϵ_{02} (19, column 10).

[7] An additional 251 cases of uncertain survival status were deleted from this illustration.

TABLE 2

SURVIVAL EXPERIENCE FOLLOWING DIAGNOSIS OF CANCER OF THE CERVIX UTERI: CASES INITIALLY DIAGNOSED 1942–1954

		Number not due for withdrawal in interval $(x, x + 1)$*				Number due for withdrawal in interval $(x, x + 1)$**			
Interval since diagnosis (years)	Number living at beginning of interval $(x, x + 1)$	Number surviving the interval	Number dying in the interval			Number living at time of withdrawal	Number dying before withdrawal		
			Total	Cancer of the cervix	Other causes		Total	Cancer of the cervix	Other causes
$x - x + 1$	N_x	s_x	δ_x	δ_{x1}	δ_{x2}	w_x	ϵ_x	ϵ_{x1}	ϵ_{x2}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0–1	5982	4030	1287	1105	182	576	89	70	19
1–2	4030	2845	644	557	87	501	40	31	9
2–3	2845	2117	250	206	44	459	19	15	4
3–4	2117	1573	151	113	38	379	14	8	6
4–5	1573	1176	87	61	26	306	4	2	2
5–6	1176	861	57	24	33	254	4	3	1
6–7	861	660	32	16	16	167	2	2	0
7–8	660	474	22	11	11	161	3	2	1
8–9	474	344	12	5	7	116	2	1	1
9–10	344	245	11	7	4	85	3	2	1
10–11	245	158	6	4	2	78	3	1	2
11–12	158	72	4	1	3	80	2	1	1
12–13	72	0	0	0	0	72	0	0	0

*Patients admitted more than $x + 1$ years prior to closing date.

**Patients admitted between x and $x + 1$ years prior to closing date.

Source: California Tumor Registry, Department of Public Health, State of California.

TABLE 3

Survival Experience After Diagnosis of Cancer of the Cervix Uteri: The Main Life Table Functions and Their Standard Errors

Interval since diagnosis (years)	x-year survival rate $\hat{p}_{0x}$		Estimated probability of death in interval $(x, x+1)$		Observed expectation of life at x	
$x - x + 1$	1000 $\hat{p}_{0x}$	1000 $S_{\hat{p}_{0x}}$	1000 $\hat{q}_x$	1000 $S_{\hat{q}_x}$	$\hat{e}_x$	$S_{\hat{e}_x}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0–1	1000.00	0.00	242.54	5.69	12.90	2.83
1–2	757.46	5.80	181.43	6.26	15.86	3.74
2–3	620.03	6.65	103.03	5.95	18.27	4.57
3–4	556.15	7.01	85.76	6.38	19.31	5.09
4–5	508.46	7.33	64.13	6.50	20.08	5.56
5–6	475.85	7.61	58.20	7.23	20.42	5.94
6–7	448.15	7.95	43.76	7.34	20.65	6.31
7–8	428.54	8.29	43.20	8.45	20.57	6.60
8–9	410.03	8.71	33.69	8.85	20.48	6.89
9–10	396.22	9.17	46.55	12.15	20.17	7.13
10–11	377.77	9.98	43.85	14.30	20.13	7.47
11–12	361.21	10.97	51.06	20.30	20.03	7.81
12–13	342.77	12.73	00.00	00.00	20.08	7.79
13	342.77	12.73	—	—	19.08	7.79

Source: California Tumor Registry, Department of Public Health, State of California.

The second interval began with the 4030 survivors from the first interval, which is entered as N_1 in line 2 of column 2. All 1953 admissions included in N_1 were due for withdrawal in the second interval.

The main life table functions and the corresponding sample standard errors as shown in Table 3 are determined from the data given in Table 2. The x-year survival rate $\hat{p}_{0x}$ is by definition equal to l_x divided by the radix l_0 , or $\hat{p}_0\hat{p}_1$, $\cdots$, $\hat{p}_{x-1}$. The sample variance of $\hat{p}_{0x}$ is computed from a formula given in a previous publication [6] (see also [13]):

$$S^2_{\hat{p}_{0x}} = \hat{p}^2_{0x} \sum_{u=0}^{x-1} \hat{p}_u^{-2} S^2_{\hat{q}_u} .$$

Formulas (13) and (14) were both used to compute the sample standard error of $\hat{q}_x$, with numerical results that were almost identical to the fourth decimal place. The figures appearing in column 5 of Table 3 were obtained by formula (14). The observed expectation of life was determined from formula (17), for which $\hat{p}_T$ was set equal to $\hat{p}_{11}$.

TABLE 4

SURVIVAL EXPERIENCE AFTER DIAGNOSIS OF CANCER OF THE CERVIX UTERI: ESTIMATED CRUDE AND NET PROBABILITIES OF DEATH FROM CANCER OF THE CERVIX UTERI AND FROM OTHER CAUSES

		Estimated crude probabilities of death in interval $(x, x + 1)$ from		Estimated net probabilities of death in interval $(x, x + 1)$ when	
Interval since diagnosis (years)	Estimated probability of surviving interval $(x, x + 1)$	Cervix cancer	Other causes	Cervix cancer Acting alone	Cervix cancer Eliminated
$x - x + 1$	$1000\ \hat{p}_x$	$1000\ \hat{Q}_{x1}$	$1000\ \hat{Q}_{x2}$	$1000\ \hat{q}_{x1}$	$1000\ \hat{q}_{x2}$
(1)	(2)	(3)	(4)	(5)	(6)
0–1	757.46	207.11	35.43	211.17	39.77
1–2	818.57	155.97	25.46	158.11	27.71
2–3	896.97	84.65	18.38	85.46	19.22
3–4	914.24	62.89	22.87	63.63	23.63
4–5	835.87	44.40	19.73	44.85	20.19
5–6	941.80	25.76	32.44	26.19	32.87
6–7	956.24	23.17	20.59	23.41	20.84
7–8	956.80	22.47	20.73	22.70	20.97
8–9	966.31	14.44	19.25	14.58	19.39
9–10	953.45	29.93	16.62	30.18	16.88
10–11	956.15	24.36	19.49	24.60	19.73
11–12	948.94	17.02	34.04	17.32	34.34
12–13	1000.00	—	—	—	—

Source: California Tumor Registry, Department of Public Health, State of California.

Table 4 shows the estimated probability of surviving each interval and the estimated crude and net probabilities of death from R_1, cancer of the cervix uteri and R_2, all other causes of death. Since only two risks are studied, the probability q_{x2} is equal to $q_{x.1}$, the net probability of death when cancer of the cervix uteri is eliminated as a risk of death from the population. For each interval the sum of $\hat{p}_x$, $\hat{Q}_{x1}$, and $\hat{Q}_{x2}$ is unity, and the estimated net probability $\hat{Q}_{xk}$ is always greater than the corresponding crude probability $\hat{Q}_{xk}$.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to George Linden and his staff at the California Tumor Registry, Department of Public

Health, State of California, for their cooperation and assistance in making their follow-up data available for our use. My thanks are also due to William F. Taylor for reading the paper.

REFERENCES

[1] Berkson, J. and Gage, R. P. [1950]. Calculation of survival rates for cancer. *Proc. Staff Meetings Mayo Clinic 25*, 270–86.

[2] Berkson, J. and Gage, R. P. [1952]. Survival curve for cancer patients following treatment. *J. Amer. Stat. Assoc. 47*, 501–15.

[3] Boag, J. W. [1949]. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Stat. Soc. 11*, 15–53.

[4] Chiang, C. L. [1958]. An application of stochastic processes to the life table and standard error of age-adjusted rates (abstract). *Biometrics 14*, 133–4.

[5] Chiang, C. L. [1960]. A stochastic study of the life table and its applications: I. Probability distributions of the biometric functions. *Biometrics 16*, 618–35.

[6] Chiang, C. L. [1960]. A stochastic study of the life table and its applications: II. Sample variance of the observed expectation of life and other biometric functions. *Human Biology 32*, 221–38.

[7] Chiang, C. L. (unpublished). On the probability of death from specific causes in the presence of competing risks. (To be published in the *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*).

[8] Dorn, H. [1950]. Methods of analysis for follow-up studies. *Human Biology 22*, 238–48.

[9] Elveback, L. [1958]. Estimation of survivorship in chronic disease: the "Actuarial" method. *J. Amer. Stat. Assoc. 53*, 420–40.

[10] Epstein, B. and Sobel, M. [1953]. Life testing. *J. Amer. Stat. Assoc. 48*, 486–502.

[11] Fix, E. [1951]. Practical implications of certain stochastic models on different methods of follow-up studies. (Presented at annual meeting of the Western Branch, A. P. H. A., Nov., 1951).

[12] Fix, E. and Neyman, J. [1951]. A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology 23*, 205–41.

[13] Greenwood, M. [1926]. A report on the natural duration of cancer. *Reports on Public Health and Medical Subjects, No. 33*, 1–26. His Majesty's Stationary Office.

[14] Harris, T. E., Meier, P. and Tukey, J. W. [1950]. Timing of the distribution of events between observations. *Human Biology 22*, 249–70.

[15] Jordcn, C. W. [1952]. *Life Contingencies*. Society of Actuaries, Chicago.

[16] Kaplan, E. L. and Meier, P. [1958]. Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc. 53*, 457–81.

[17] Karn, M. N. [1933]. A further study of methods of constructing life tables when certain causes of death are eliminated. *Biometrika 25*, 91.

[18] Littell, A. S. [1952]. Estimation of the T-year survival rate from follow-up studies over a limited period of time. *Human Biology 24*, 87–116.

[19] Neyman, J. [1950]. *First course in probability and statistics*. Henry Holt, New York.

THE SPEARMAN ESTIMATOR FOR SERIAL DILUTION ASSAYS[1]

Eugene A. Johnson and Byron Wm. Brown, Jr.
School of Public Health, University of Minnesota,
Minneapolis 14, Minnesota, U.S.A.

SUMMARY

A new estimation procedure is presented for the estimation of the density of organisms in a suspension by a serial dilution assay. The point estimator is based on the Spearman technique, is simply and quickly computed without special tables, and has efficiency 88 percent.

1. INTRODUCTION

To estimate the density of a specific organism in a suspension, a common method is to form a series of dilutions of the original suspension. Then from each dilution a specified volume, hereafter referred to as a dose, is placed in each of several tubes. Later the tubes are examined for evidence of growth of the organism.

Suppose $k + 1$ dilutions are used with concentrations of $z_i = a^{-i}z_0$, $i = 0, +1, +2, \cdots, +k$. z_0 is the highest concentration of original suspension used and $a > 1$ is the dilution factor.

The following notation will be used:

$x = \ln z$,
n = the number of tubes receiving a unit volume (dose) for each dilution,
r_i = the number of tubes showing signs of growth at dilution z_i,
$p_i = r_i/n$, $s_i = n - r_i$, $q_i = 1 - p_i$.

The usual model specifies a density of organisms of θ per unit volume in the original suspension and a Poisson distribution of organisms in the individual doses. Thus the probability of signs of growth (i.e., the probability of one or more organisms) in a tube receiving a dose at concentration z_i is

$$P(z_i) = 1 - e^{-\theta z_i}. \qquad (1)$$

The maximum likelihood estimator (frequently called "most probable number") has been proposed [5] for estimating θ. Computational

[1] Research supported, in part, by a National Institute of Health Research Training Grant.

procedures and tables have been published to eliminate the tedious computations involved [1, 4, 7]. Cochran [2] has given some guides for designing serial dilution assays using the most probable number.

Fisher [5] proposed another estimator for θ. Computation is tedious for this estimator also. Tables have been published [6] for some of the common experimental designs.

2. POINT AND INTERVAL ESTIMATION PROCEDURES

The dose levels on the log scale, $x = \ln z$, are spaced $d = \ln a$ units apart. The probability of a growth response at dose level x is

$$F(x) = 1 - e^{-\theta e^{x}}. \tag{2}$$

The equidistant spacing and the "dose-response" function $F(x)$ with the characteristics of a distribution function (increasing from 0 to 1 with x) are typical of the common bioassay problem in which the object is to estimate the median (L.D. 50) or mean (μ) or $F(x)$. For such situations, Spearman [9] proposed the following estimator for μ:

$$\hat{\mu} = x_0 + \frac{d}{2} - d \sum p_i \ , \tag{3}$$

where $x_0 = \ln z_0$, $d = \ln a$, the summation is from $i = 0$ to $i = k$, and p_i is the proportion of tubes showing growth at dilution z_i . $F(x)$ has mean:

$$\begin{aligned} \mu &= \int_{-\infty}^{\infty} x\theta e^{x} e^{-\theta e^{x}}\, dx \\ &= \int_{0}^{\infty} (\ln y - \ln \theta) e^{-y}\, dy, \\ &= -\gamma - \ln \theta, \end{aligned}$$

where $\gamma = .57722$ is Euler's constant [8]. The parameter of interest θ can then be written:

$$\theta = e^{-\gamma} e^{-\mu} \tag{4}$$

and the point estimator $\hat{\theta}$ of θ based on $\hat{\mu}$ is[2]

$$\hat{\theta} = e^{-\gamma} e^{-\hat{\mu}} = e^{-\gamma - x_0 - d/2 + d\Sigma p_i}. \tag{5}$$

The point estimator of the number of organisms per z_0 volume is

$$z_0 \hat{\theta} = e^{-\gamma - d/2 + d\Sigma p_i}. \tag{6}$$

[2] The point estimator $\hat{\theta}$ is biased. Its bias is discussed in Section 4 and a correction factor is given.

For convenience, the necessity of a logarithm table can be circumvented by constructing a graph of $z_0\hat{\theta}$ as a function of $\sum p_i$, for commonly used values of a. Such a graph has been constructed (Graph 1) and its use is illustrated in the example in Section 6.

If an interval estimate of θ is desired, the interval estimate of μ is computed first and then the end points transformed by equation (5). The interval estimator of μ (confidence coefficient of .95) is given in expression (7). The justification is presented in Section 3.

$$\hat{\mu} - 1.96\sqrt{\frac{d \ln 2}{n}} < \mu < \hat{\mu} + 1.96\sqrt{\frac{d \ln 2}{n}}. \tag{7}$$

The corresponding interval estimator (C.C. = .95) for θ is

$$\hat{\theta} \exp\left(-1.96\sqrt{\frac{d \ln 2}{n}}\right) < \theta < \hat{\theta} \exp\left(1.96\sqrt{\frac{d \ln 2}{n}}\right), \tag{8}$$

where $\hat{\theta}$ is given in equation (5).

3. CHARACTERISTICS OF $\hat{\mu}$

First the values of d, n and x_0 will be assumed fixed. Then the case of x_0 randomly chosen will be discussed. In all cases it is assumed that the dose levels $x_i = \ln z_i$ span an interval wide enough so that $F(x_0) \geq .99$ and $F(x_0 - kd) \leq .01$.

The following notation will be used to differentiate between discussions with x_0 fixed and discussions with x_0 random:

E_p denotes expectation with respect to the random variables p_i for fixed x_0.

E_{x_0} denotes expectation with respect to the distribution of x_0 when x_0 is randomly chosen.

E denotes expectation with respect to both x_0 and the random variables p_i.

It follows from (2) and (3) that the expected value of $\hat{\mu}$ for fixed x_0 is

$$\begin{aligned} E_p(\hat{\mu} \mid x_0) &= x_0 + \frac{d}{2} - d \sum F(x_i) \\ &= x_0 + \frac{d}{2} - d \sum (1 - \exp[-\theta e^{x_0 - id}]). \end{aligned} \tag{9}$$

Using (4) and (9), the bias of $\hat{\mu}$ for fixed x_0, $B(\hat{\mu} \mid x_0)$, is

$$\begin{aligned} B(\hat{\mu} \mid x_0) &= E_p(\hat{\mu} \mid x_0) - \mu \\ &= x_0 + \frac{d}{2} - d \sum (1 - \exp[-\theta e^{x_0 - id}]) + \gamma + \ln \theta. \end{aligned} \tag{10}$$

$B(\hat{\mu} \mid x_0)$ does not depend on n. It does depend on d and on x_0. For a given value of d, the bias will have a maximum and a minimum over all possible positions of x_0, subject to the conditions that $F(x_0) \geq .99$ and $F(x_0 - kd) \leq .01$. Let $B(\hat{\mu} \mid x_0)$ be differentiated with respect to x_0:

$$\frac{d}{dx_0}[B(\hat{\mu} \mid x_0)] = 1 - d \sum \theta \exp(x_0 - id) \exp(-\theta e^{x_0 - id}). \tag{11}$$

For a given d, if enough terms are carried to cover the desired range, it can be shown that (11) has two zeros with respect to x_0 on every interval of length d. These solutions yield the maximum and minimum values of the bias over all positions x_0 for fixed d. Some results, obtained numerically, are:

$$\begin{aligned} a &= 4, & d &= \ln 4: & |\, B(\hat{\mu} \mid x_0) \,| &\leqq .0019, \\ a &= 10, & d &= \ln 10: & |\, B(\hat{\mu} \mid x_0) \,| &\leqq .0417. \end{aligned} \tag{12}$$

Using (2) and (3), the variance of $\hat{\mu}$ for fixed x_0, $V(\hat{\mu} \mid x_0)$, can be written

$$V(\hat{\mu} \mid x_0) = \frac{d^2}{n} \sum F(x_i)[1 - F(x_i)]. \tag{13}$$

The sum in (13) can be approximated by the following integral. The approximation will be good if d is small and the range of dose levels x_i is large.

$$V(\hat{\mu} \mid x_0) \doteq \frac{d}{n} \int_{-\infty}^{\infty} F(x)[1 - F(x)]\, dx = \frac{d \ln 2}{n}. \tag{14}$$

For a given value of d, $V(\hat{\mu} \mid x_0)$ will depend on the position of x_0. The maximum values of $V(\hat{\mu} \mid x_0)$ can be obtained for given d by differentiating (13):

$$\frac{d}{dx_0}[V(\hat{\mu} \mid x_0)]$$
$$= \frac{d^2}{n} \sum \theta e^{x_0 - id} \exp(-\theta e^{x_0 - id})[2 \exp(-\theta e^{x_0 - id}) - 1]. \tag{15}$$

Zeros for (15) can be obtained numerically for any given d, using a range of doses large enough so that terms ignored in (15) are negligible. The values of x_0 thus obtained can be substituted in (13) to obtain the maximum and minimum variance. The results below are presented relative to the approximate variance (14):

$$\begin{aligned} a = 4, \qquad d = \ln 4&: \quad \left| \frac{V(\hat{\mu} \mid x_0) - \dfrac{d \ln 2}{n}}{\dfrac{d \ln 2}{n}} \right| \leqq .0056, \\ a = 10, \qquad d = \ln 10&: \quad \left| \frac{V(\hat{\mu} \mid x_0) - \dfrac{d \ln 2}{n}}{\dfrac{d \ln 2}{n}} \right| \leqq .0977. \end{aligned} \tag{16}$$

The numerical results (12) and (16) indicate that, if the range of doses is wide enough and the dilution factor a is 10 or less, then the following approximations can be used regardless of the position of x_0 :

$$E_p(\hat{\mu} \mid x_0) \doteq \mu, \qquad V(\hat{\mu} \mid x_0) \doteq \frac{d \ln 2}{n}. \tag{17}$$

Suppose that x_0 is randomly chosen with uniform density over an interval of length d, say $(A, A + d)$. Then the expectation of μ can be taken with respect to x_0 and the p_i :

$$\begin{aligned} E(\hat{\mu}) = E_{x_0}[E_p(\hat{\mu} \mid x_0)] &= \frac{1}{d} \int_A^{A+d} E_p(\hat{\mu} \mid x_0) \, dx_0 \\ &= \frac{1}{d} \int_A^{A+d} \left[x_0 + \frac{d}{2} - d \sum F(x_i) \right] dx_0 \\ &= (A + d) - \sum \int_A^{A+d} F(x_0 - id) \, dx_0 \, . \\ &= (A + d) - \int_{A-kd}^{A+d} F(x) \, dx. \end{aligned} \tag{18}$$

Integrating (18) by parts:

$$E(\hat{\mu}) = \int_{A-kd}^{A+d} xf(x) \, dx. \tag{19}$$

If the doses extend over a wide enough range, (19) indicates that $\hat{\mu}$ has a negligible bias:

$$E(\hat{\mu}) \doteq \mu. \tag{20}$$

For x_0 randomly chosen on $(A, A + d)$ the variance of $\hat{\mu}$ can be written:

$$\begin{aligned} V(\hat{\mu}) = E(\hat{\mu} - \mu)^2 &= E_{x_0}[V(\hat{\mu} \mid x_0) + B^2(\hat{\mu} \mid x_0)] \\ &= \frac{1}{d} \int_A^{A+d} \frac{d^2}{n} \sum F(x_i)[1 - F(x_i)] \, dx_0 + \frac{1}{d} \int_A^{A+d} B^2(\hat{\mu} \mid x_0) \, dx_0 \, . \end{aligned}$$

$$V(\hat{\mu}) = \frac{d}{n} \int_{A-kd}^{A+d} F(x)[1 - F(x)] \, dx + \frac{1}{d} \int_A^{A+d} B^2(\hat{\mu} \mid x_0) \, dx_0 \, . \tag{21}$$

If the doses extend over a range such that the first integral in (21) is approximated to the degree desired by integrating over the infinite range, $V(\hat{\mu})$ can be written

$$V(\hat{\mu}) = \frac{d}{n}\int_{-\infty}^{\infty} F(x)[1 - F(x)]\,dx + \frac{1}{d}\int_{A}^{A+d} B^2(\hat{\mu} \mid x_0)\,dx_0 . \tag{22}$$

The first integral has the value $(d \ln 2)/n$. It follows from (22) that

$$\frac{\left| V(\hat{\mu}) - \dfrac{d \ln 2}{n} \right|}{\dfrac{d \ln 2}{n}} \leqq \frac{n}{d \ln 2}\,[\max_{x_0} B^2(\hat{\mu} \mid x_0)]. \tag{23}$$

From (23) and the numerical results in (12) the following bounds for the approximation of $V(\hat{\mu})$ by $(d \ln 2)/n$ can by obtained:

$$\begin{aligned} a &= 4, \qquad d = \ln 4: \quad \frac{\left| V(\hat{\mu}) - \dfrac{d \ln 2}{n} \right|}{\dfrac{d \ln 2}{n}} \leqq (3.76n)10^{-6}, \\ a &= 10, \qquad d = \ln 10: \quad \frac{\left| V(\hat{\mu}) - \dfrac{d \ln 2}{n} \right|}{\dfrac{d \ln 2}{n}} \leqq (1.09n)10^{-3}. \end{aligned} \tag{24}$$

A summary of the results (20) and (24), for randomly chosen x_0, would be essentially the same as (17), i.e., if the doses extend over a wide enough range,

$$E(\hat{\mu}) \doteq \mu, \qquad V(\hat{\mu}) \doteq \frac{d \ln 2}{n}. \tag{25}$$

In concluding the remarks on the properties of $\hat{\mu}$, it should be remarked that the distribution of $\hat{\mu}$, conditional on x_0, is asymptotically normal, since it is a linear function (3) of the sum of $k + 1$ binomially distributed random variables.

4. ESTIMATION OF θ

The point estimator of θ based on $\hat{\mu}$ was defined in (5). The expected value of this estimator can be approximated for random choice of x_0 as follows:

$$\begin{aligned} E(\hat{\theta}) &= E[e^{-\gamma-\hat{\mu}}] = E[e^{-(\gamma+\mu)-(\hat{\mu}-\mu)}] \\ &= \theta E[e^{-(\hat{\mu}-\mu)}] = \theta E\left[\sum_{i=0}^{\infty} \frac{(\hat{\mu} - \mu)^i}{i!}\right]. \end{aligned} \tag{26}$$

If the expectation is taken termwise in (26), terms in moments of order three and higher are dropped, and the approximation (25) is used, (26) can be written:

$$E(\hat{\theta}) \doteq \theta\left[1 + \frac{d \ln 2}{2n}\right]. \tag{27}$$

From (27) an estimator that is less biased than $\hat{\theta}$ is

$$\hat{\theta}' = \frac{2n}{2n + d \ln 2}\,\hat{\theta}. \tag{28}$$

The coefficient of variation of $\hat{\theta}$, C.V. $(\hat{\theta})$, can be approximated using (5) and (24):

$$C.V.(\hat{\theta}) = \frac{\sqrt{V(\hat{\theta})}}{\theta} \doteq \frac{\sqrt{\theta^2 V(\hat{\mu})}}{\theta} \doteq \sqrt{\frac{d \ln 2}{n}}. \tag{29}$$

The approximation to the variance of $\hat{\mu}$, given in (17) and in (25), is identical with the asymptotic variance of the estimator for ln θ proposed by Fisher [5]. Therefore the asymptotic efficiency of the two procedures for estimating θ will be the same, 88 percent.

The interval estimator of θ given in (8) follows directly from (4), (17) and (25), and the asymptotic normality of $\hat{\mu}$.

5. DESIGN OF THE SERIAL DILUTION ASSAY

In choosing the range of dilutions, the conditions $F(x_0) \geq .99$ and $F(x_0 - kd) \leq .01$ should be satisfied. This implies that the expected number of organisms per dose at the highest and lowest concentrations should be greater than 5 and less than .01 respectively.

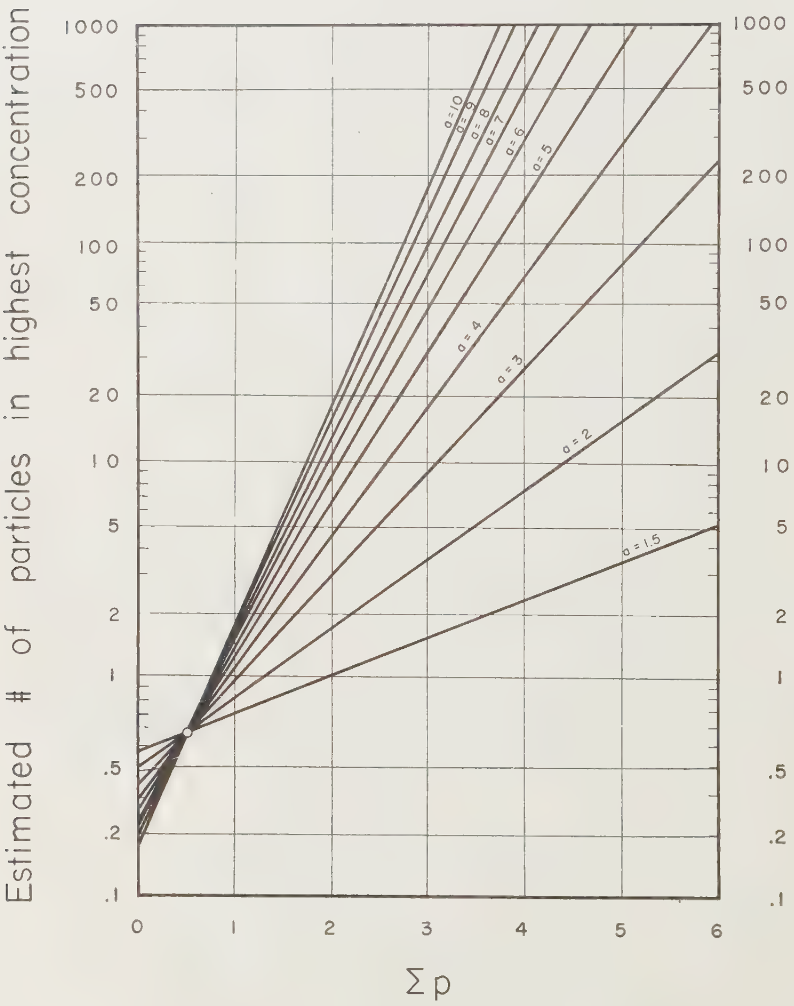
In designing a serial dilution assay the precision desired of the estimator must be specified. This can be done by specifying the desired value of the coefficient of variation (29). Then equation (29) can be used to obtain n, the number of tubes per dilution, for any dilution constant a desired.

Alternatively the precision of the estimator can be specified in terms of the interval estimator of θ. Using (8), the precision of the interval estimator for θ can be expressed in terms of the factor, $R = e^{1.96\sqrt{d\ln 2/n}} > 1$, which will be multiplied and divided into $\hat{\theta}$ to obtain the upper and lower limits of the 95 percent confidence interval. Graph 2 shows the values of R as functions of the dilution factor a for values of n from one to ten. This graph can be used to design dilution assays having desired precision. See the example in Section 6.

6. EXAMPLE OF DESIGN AND ANALYSIS

Suppose it is known that the density of organisms θ is between 10^2 and 10^5 organisms per unit volume. Suppose a ten-fold dilution factor is to be used. To assure doses spanning the range .01 to 5 organisms per dose, it is necessary to use concentrations ranging from 10^{-1} to 10^{-7} dilutions of the original suspension, or 7 dilutions altogether.

Suppose the value of θ is desired to within approximately four fold

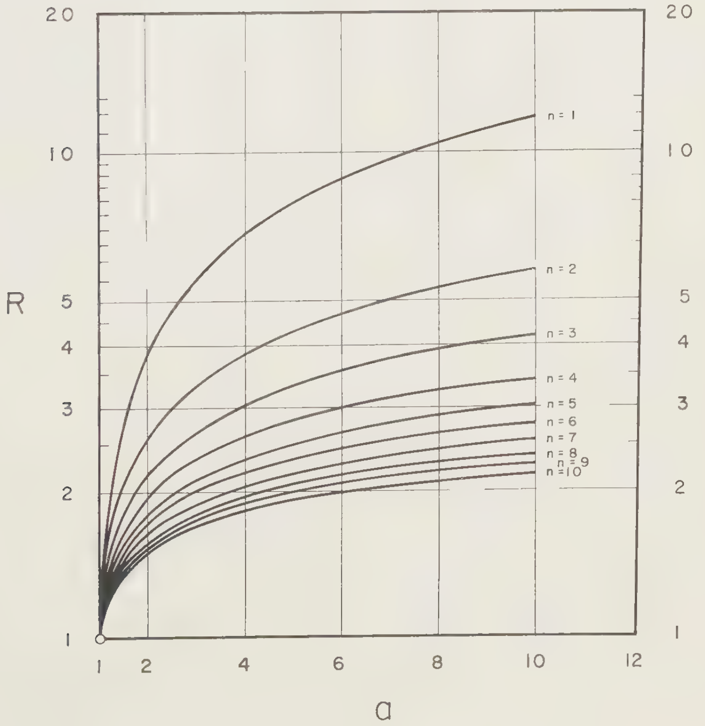


GRAPH 1

DENSITY ESTIMATES AS A FUNCTION OF $\sum p$ AND a

TABLE I
Illustrative Data

Dilution	Proportion of Tubes Showing Growth p
10^{-1}	3/3
10^{-2}	3/3
10^{-3}	2/3
10^{-4}	0/3
10^{-5}	0/3
10^{-6}	0/3
10^{-7}	0/3
	$\sum p = 8/3$



GRAPH 2
The 95 Per Cent Confidence Factor (R) for Combinations of n and a

in either direction (i.e., the desired value of R is approximately 4). From Graph 2, $n = 3$ tubes per dilution will give $R = 4.2$ for a dilution factor of 10. This means a total sample size of $7n = 21$ tubes. In Table 1 some data and computational results are presented for such a design.

Using Graph 1, $z_0\hat{\theta} \doteq 80$ organisms per dose at the highest concentration, 10^{-1}. Therefore, $\hat{\theta} \doteq 800$ organisms per unit volume of the original suspension. Without Graph 1, $\hat{\theta}$ would be computed as follows:

$$\hat{\mu} = \ln 10^{-1} + \frac{\ln 10}{2} - (\ln 10)\left(\frac{8}{3}\right) = -7.2915,$$

$\hat{\theta} = e^{-\gamma-\hat{\mu}} = 824$ organisms per unit volume of the original suspension.

The interval estimate of $\hat{\theta}$ is obtained by using the factor R $(= 4.2)$ read from Graph 2:

$$\frac{824}{4.2} < \theta < 824(4.2),$$

$$196 < \theta < 3460 \text{ (confidence coefficient of .95).}$$

The point estimate $\hat{\theta}$ can be corrected for bias using equation (28):

$$\hat{\theta}' = \frac{2n}{2n + d \ln 2}\,\hat{\theta} = (.79)(824) = 651 \text{ organisms per unit volume.}$$

REFERENCES

[1] Barkworth, H. and Irwin, J. O. [1938]. Distribution of Coliform Organisms in Milk and the Accuracy of the Presumptive Coliform Test. *Jour. Hygiene 38*, 446–57.

[2] Cochran, W. G. [1950]. Estimation of Bacterial Densities by Means of the "Most Probable Number". *Biometrics 6*, 105–16.

[3] Finney, D. J. [1947]. The Principles of Biological Assay. *Jour. Roy. Stat. Soc., Suppl. 9.* 46–91.

[4] Finney, D. J. [1952]. *Statistical Methods in Biological Assay*, Hafner, New York.

[5] Fisher, R. A. [1922]. On the Mathematical Foundations of Theoretical Statistics. *Phil. Trans. Royal Soc. A222*, 309–68.

[6] Fisher, R. A. and Yates, F. [1957]. *Statistical Tables for Biological Agricultural and Medical Research,* (5th edition). Edinburgh: Oliver and Boyd.

[7] Halvorson, H. O. and Ziegler, N. R. [1933]. Applications of Statistics to Problems in Bacteriology. I. A means of determining bacterial population by the dilution method. *Jour. Bacteriology 25*, 101–21.

[8] Ryshik, I. M. and Gradstein, I. S. [1957]. Tables of Series, Products and Integrals. Veb. Deutscher Verlag der Wissenschaften, Berlin, 306.

[9] Spearman, C. [1908]. The Method of 'Right and Wrong Cases' ('Constant Stimuli') without Gauss's Formulae. *Brit. Jour. Psych. 2*, 227–42.

THE FITTING OF A GENERALIZATION OF THE LOGISTIC CURVE

J. A. Nelder
National Vegetable Research Station, Wellesbourne, England

This paper describes the fitting of the four-parameter family of curves defined by the differential equation.

$$\frac{dW}{dt} = \kappa W\left[1 - \left(\frac{W}{A}\right)^{1/\theta}\right] \tag{1}$$

Special cases were originally proposed by Pütter [1920] for various types of animal growth (see e.g. von Bertalanffy [1957]), and recently Richards [1959] has exemplified the general form of the curves and suggested that they may be useful for the empirical description of plant growth. For further details of the history of these curves and their mathematical properties reference should be made to Richards' paper. It suffices to say here that the family defined by (1) includes as special cases several curves which have been used empirically for the description of growth, including the 'monomolecular' (diminishing returns) curve ($\theta = -1$), the exponential curve ($\theta \rightarrow 0$ through positive values), the logistic curve ($\theta = 1$), and the Gompertz curve ($\theta \rightarrow \infty$ with A fixed and κ a linear function of θ).

MODEL AND NOTATION

In using (1) in the description of growth, W will usually denote a weight of some kind, while t is usually chronological time, but may be a suitable 'time scale' (Skellam *et al.* [1959], Nelder *et al.* [1960]) which can replace chronological time when the environment is variable. Since it seems to be characteristic of many growth phenomena that the relative growth rate ($W^{-1}\, dW/dt$) is almost constant when the weight is small compared to the final weight, we shall restrict θ to be positive in what follows. For if $\theta < 0$, then, for W sufficiently small, $W^{-1}\, dW/dt$ becomes as large as we please in absolute value. When $\theta > 0$, the solution of (1) can be written in the form

$$W = A/\{1 + e^{-(\lambda+\kappa t)/\theta}\}^{\theta} \tag{2}$$

where λ is the constant of integration. A and κ are taken as positive since W is positive and is assumed to increase with time. Thus $W^{1/\theta}$ satisfies the logistic equation with upper asymptote $A^{1/\theta}$ as $t \geq \infty$, while for t large and negative

$$W \sim Ae^{\lambda+\kappa t}.$$

We now write $Y = \ln W$ and $\alpha = \ln A$ so that

$$Y = \alpha - \theta \ln [1 + e^{(\lambda+\kappa t)/\theta}] \tag{3}$$

where ln denotes the natural logarithm. (3) is the form of the equation that will be used for fitting, and we shall assume that if $y = \ln w$ is a sample value of $Y = \ln W$ at time t, then $E(y) = Y$, and var $y = \sigma^2$ independently of Y. We shall also assume that the y_i $(i = 1, \cdots, n)$ obtained at time t_i, i.e. the sample points to which the curve is to be fitted, are independent. The independence condition is satisfied for instance in plant growth analysis when a randomized field layout is used, and the sampling is destructive, i.e., a different set of plants is taken on each occasion. When the same set of plants is used on each occasion (when w might denote leaf area measured non-destructively) it is still possible to use the method proposed for estimation, though of course sampling variances cannot be obtained from residual mean squares 'within curves' if the condition of independence is not satisfied. This situation is further discussed later in the paper.

The assumption that $y = \ln w$ has constant variance has been found to be reasonable from an examination of data on the weight of part or the whole of several crops. In fact it is usually found that the distribution of y does not differ significantly from a normal distribution.

For the purposes of this paper, (2) will be described as the *generalized logistic equation*, though it is not of course the only generalization that could or has been made from the logistic equation.

FITTING THE LOGISTIC CURVE

We consider first the case where θ is known. This is equivalent to fitting the logistic equation, since by putting $W' = W^{1/\theta}$, $Y' = Y/\theta$, $\alpha' = \alpha/\theta$, $\lambda' = \lambda/\theta$ and $\kappa' = \kappa/\theta$ we can convert (2) into the logistic function. We may, therefore, put $\theta = 1$ for this case without loss of generality. Let us write

$$\tau_i = \lambda + \kappa t_i \quad \text{and} \quad \xi_i = \frac{e^{-\tau_i}}{1 + e^{-\tau_i}} = \frac{1}{1 + e^{\tau_i}} = \frac{A - W_i}{A};$$

then from (3)

$$Y_i = \alpha + \ln(1 - \xi_i), \quad \frac{\partial Y_i}{\partial \alpha} = 1, \quad \frac{\partial Y_i}{\partial \lambda} = \xi_i\ , \quad \text{and} \quad \frac{\partial Y_i}{\partial \kappa} = \xi_i t_i\ ,$$

Thus the least-square equations are given by

$$\begin{aligned} \sum_i (y_i - Y_i) &= 0, \\ \sum_i (y_i - Y_i)\xi_i &= 0, \\ \sum_i (y_i - Y_i)\xi_i t_i &= 0. \end{aligned} \tag{4}$$

These equations have no explicit solution in general, so that they must be solved by iteration. The usual method, using the expected values of the information matrix (see e.g. Bailey [1951]), gives the following

TABLE I

Extension of Berkson's Table of Anti-Logits $[1/1 + e^{-\tau})]$

τ	0	1	2	3	4	5	6	7	8	9
					0.99					
5.0	331	337	344	350	357	363	369	376	382	388
5.1	394	400	406	412	418	423	429	435	440	446
5.2	451	457	462	467	473	478	483	488	493	498
5.3	503	508	513	518	523	527	532	537	541	546
5.4	550	555	559	564	568	572	576	581	585	589
5.5	593	597	601	605	609	613	617	620	624	628
5.6	632	635	639	642	646	649	653	656	660	663
5.7	667	670	673	676	680	683	686	689	692	695
5.8	698	701	704	707	710	713	716	719	721	724
5.9	727	730	732	735	737	740	743	745	748	750
6.0	753	755	758	760	762	765	767	769	772	774
6.1	776	778	781	783	785	787	789	791	793	795
6.2	797	799	801	803	805	807	809	811	813	815
6.3	817	819	820	822	824	826	827	829	831	832
6.4	834	836	837	839	841	842	844	845	847	848
6.5	850	851	853	854	856	857	859	860	861	863
6..6	864	865	867	868	869	871	872	873	875	876
6.7	877	878	879	881	882	883	884	885	887	888
6.8	889	890	891	892	893	894	895	896	897	898
6.9	899	900	901	902	903	904	905	906	907	908

To obtain ξ, from this table and Berkson's table in Berkson [1953], use the formula, ξ = antilogit $(-\tau)$ = 1 − antilogit (τ).

TABLE II

Values of $10^4 \ln (1 + e^{-\tau})$

For τ positive $-\ln (1 - \xi) =$ tabulated value.
For τ negative $-\ln (1 - \xi) = |\tau| +$ tabulated value for $|\tau|$.

τ	0	1	2	3	4	5	6	7	8	9
0.0	6931	6882	6832	6783	6733	6685	6636	6588	6539	6492
0.1	6444	6397	6349	6303	6256	6210	6163	6118	6072	6027
0.2	5981	5936	5892	5847	5803	5759	5716	5672	5629	5586
0.3	5544	5501	5459	5417	5375	5334	5293	5252	5211	5170
0.4	5130	5090	5050	5011	4972	4932	4894	4855	4817	4779
0.5	4741	4703	4666	4629	4592	4555	4518	4482	4446	4410
0.6	4375	4340	4304	4270	4235	4201	4166	4132	4099	4065
0.7	4032	3999	3966	3933	3901	3869	3837	3805	3773	3742
0.8	3711	3680	3649	3619	3589	3559	3529	3499	3470	3441
0.9	3412	3383	3354	3326	3298	3270	3242	3214	3187	3160
1.0	3133	3106	3079	3053	3027	3001	2975	2949	2924	2898
1.1	2873	2848	2824	2799	2775	2751	2727	2703	2679	2656
1.2	2633	2610	2587	2564	2542	2519	2497	2475	2453	2432
1.3	2410	2389	2368	2347	2326	2305	2285	2264	2244	2224
1.4	2204	2184	2165	2146	2126	2107	2088	2070	2051	2032
1.5	2014	1996	1978	1960	1942	1925	1907	1890	1873	1856
1.6	1839	1822	1806	1789	1773	1757	1741	1725	1709	1693
1.7	1678	1662	1647	1632	1617	1602	1588	1573	1558	1544
1.8	1530	1516	1502	1488	1474	1460	1447	1433	1420	1407
1.9	1394	1381	1368	1355	1343	1330	1318	1306	1293	1281
2.0	1269	1257	1246	1234	1222	1211	1200	1188	1177	1166
2.1	1155	1144	1134	1123	1112	1102	1091	1081	1071	1061
2.2	1051	1041	1031	1021	1012	1002	993	983	974	965
2.3	955	946	937	928	920	911	902	894	885	877
2.4	868	860	852	844	836	828	820	812	804	796
2.5	789	781	774	766	759	752	745	737	730	723
2.6	716	710	703	696	689	683	676	670	663	657
2.7	650	644	638	632	626	620	614	608	602	596
2.8	590	585	579	573	568	562	557	551	546	541
2.9	536	530	525	520	515	510	505	500	495	491
3.0	486	481	476	472	467	463	458	454	449	445
3.1	441	436	432	428	424	420	415	411	407	403
3.2	399	396	392	388	384	380	377	373	369	366
3.3	362	359	355	352	348	345	341	338	335	331
3.4	328	325	322	319	316	312	309	306	303	300
3.5	297	295	292	289	286	283	280	278	275	272
3.6	270	267	264	262	259	257	254	252	249	247
3.7	244	242	239	237	235	232	230	228	226	223
3.8	221	219	217	215	213	211	208	206	204	202
3.9	200	198	196	195	193	191	189	187	185	183
4.0	181	180	178	176	174	173	171	169	168	166
4.1	164	163	161	160	158	156	155	153	152	150
4.2	149	147	146	144	143	142	140	139	137	136
4.3	135	133	132	131	130	128	127	126	124	123
4.4	122	121	120	118	117	116	115	114	113	112
4.5	110	109	108	107	106	105	104	103	102	101

TABLE II—(*Continued*)

τ		τ		τ		τ		τ	
4.60	100	5.10	61	5.60	37	6.10	22	6.60	14
4.70	91	5.20	55	5.70	33	6.20	20	6.70	12
4.80	82	5.30	50	5.80	30	6.30	18	6.80	11
4.90	74	5.40	45	5.90	27	6.40	17	6.90	10
5.00	67	5.50	41	6.00	25	6.50	15	—	—

adjustments $\delta\alpha_0$, $\delta\lambda_0$, $\delta\kappa_0$ to initial 'guesses' α_0 , λ_0 , κ_0 for the parameters:—

$$\begin{bmatrix} n & \sum \xi_i & \sum \xi_i t_i \\ & \sum \xi_i^2 & \sum \xi_i^2 t_i \\ & & \sum \xi_i^2 t_i^2 \end{bmatrix} \begin{bmatrix} \delta\alpha_0 \\ \delta\lambda_0 \\ \delta\kappa_0 \end{bmatrix} = \begin{bmatrix} \sum (y_i - Y_i) \\ \sum \xi_i(y_i - Y_i) \\ \sum \xi_i t_i(y_i - Y_i) \end{bmatrix} \qquad (5)$$

where ξ_i is evaluated with $\lambda = \lambda_0$ and $\kappa = \kappa_0$. New values of ξ_i and Y_i are then recalculated using $\alpha_1 = \alpha_0 + \delta\alpha_0$ etc. and further adjustments obtained, the process being repeated until whatever accuracy required is achieved. Note that the information matrix depends on the parameters only through τ_i ; the process of solution of (5) would obviously be considerably speeded up if tables of ξ_i and ln $(1 - \xi_i)$ as functions of τ_i were available.

Now ξ = antilogit $(-\tau)$ = 1 − antilogit (τ) in Berkson's [1953] notation; hence ξ can be obtained for the range $-4.99 \leq \tau \leq 4.99$ from the table of antilogits given in that paper. An extension of this table for $5 \leq |\tau| \leq 6.99$ which covers the region down to $W/A = 0.001$ and up to $W/A = 0.999$ is given in Table I. Values of ln $(1 + e^{-\tau})$ to four decimal places are given in Table II for $-6.99(0.01)6.99$ from which ln $(1 - \xi)$ can be quickly derived. These tables have been computed from the tables of e^x and ln x given in Comrie [1949] and should prove sufficiently accurate for most purposes. If linear interpolation is used from the nearest tabulated value, using the tabulated first difference, rounding-off errors will rarely exceed 1 unit in the last place.

Because Y_i is linear in α, knowledge of its approximate value is not necessary for the iterative process using the expected information matrix, and we can replace $\delta\alpha_0$ in (5) by α_1 , and Y by ln $(1 - \xi)$, solving directly for α_1 , $\delta\lambda_0$ and $\delta\kappa_0$. However when we come to consider the exact equations for the iteration, using the empirical information matrix instead of the expected one, we shall find no advantage in suppressing α_0, so that it will be retained in what follows. The calculations can be conveniently laid out as follows:

TABLE III

Computing Layout for Fitting the Logistic Curve

(1)	(2)	(3)	(4)	(5)	(6)	(7)
t_i	$\tau_i = \lambda + \kappa t_i$	ξ_i	$\xi_i t_i$	$\ln(1 - \xi_i)$	y_i	$y_i - Y_i$

The t_i are entered in column 1, whence the τ_i are calculated using the starting values of λ and κ, and entered in column 2. Column 3 is filled up from Table I, and then column 4 calculated. Column 5 is filled up from Table II, and the y's entered in column 6. α is estimated by $\bar{y} - \sum \ln(1 - \xi_i)/n$ from the sums of columns 5 and 6, and hence $y_i - Y_i = y_i - \alpha - \ln(1 - \xi_i)$ is calculated. The coefficients on the left-hand side of (5) can now be calculated from columns 3 and 4, by summing the columns and from sums of squares and products of ξ_i and $\xi_i t_i$. Finally the right-hand side of (4) may be calculated from $\sum (y_i - Y_i)$, which should be zero, apart from rounding errors, and from products of columns 3 and 4 with column 7. The equations (4) may now be solved using any of the standard techniques (see e.g., Dwyer [1951]).

The exact iterative equations

If we now denote the parameters to be fitted by θ_i, then the above method uses the expected information matrix whose general term is

$$\sum_\kappa \frac{\partial Y_\kappa}{\partial \theta_i} \frac{\partial Y_\kappa}{\partial \theta_j}.$$

The exact equations for the generalized Newton-Raphson process are those using the empirical information matrix with general term

$$\sum_\kappa \frac{\partial Y_\kappa}{\partial \theta_i} \frac{\partial Y_\kappa}{\partial \theta_j} - \sum_\kappa (y_\kappa - Y_\kappa) \frac{\partial^2 Y_\kappa}{\partial \theta_i \, \partial \theta_j}.$$

Cornfield and Mantel [1950] for the analogous situation in probit analysis found a substantial improvement in the speed of convergence using this latter form, so that it may be worth considering the appropriate modification here. This necessitates adding to the left-hand-side of (5) the matrix

$$\begin{bmatrix} 0 & 0 & 0 \\ & \sum (y_i - Y_i)\xi_i(1 - \xi_i) & \sum (y_i - Y_i)\xi_i(1 - \xi_i)t_i \\ & & \sum (y_i - Y_i)\xi_i(1 - \xi_i)t_i^2 \end{bmatrix}$$

and two more columns need to be added to Table III, namely $(y - Y)(1 - \xi)$ and ξt^2 in order to compute the elements of this matrix. The extra calculation is not large but the increase in speed of convergence achieved may justify the extra labour only when σ^2 is large so that the $(y_i - Y_i)$ are appreciable.

The dispersion matrix of the estimates of the parameters may be obtained in the usual way by inverting the information matrix on the left-hand side of (5) and then multiplying it by an estimate of σ^2. For methods of matrix inversion, and the combination of the inversion with the solution of an associated set of equations, see e.g., Dwyer [1951].

If a variance for $\hat{A}$ rather than $\hat{\alpha}$ is wanted, then we may use the approximate formula

$$\text{var } \hat{A} = \text{var } e^{\hat{\alpha}} \sim e^{2\hat{\alpha}} \text{ var } \hat{\alpha} = \hat{A}^2 \text{ var } \hat{\alpha}$$

provided that var $\hat{A}$ is small compared to $\hat{A}^2$.

Obtaining the starting values

The starting values may be obtained in various ways, some wholly graphical and some partly so. Some compromise must be made between the desirability of having them as accurate as possible so that iterations are reduced to a minimum, and the undesirability of spending an excessive amount of time on a preliminary stage of the computations.

A semi-graphical method which I have found useful is as follows. First plot the logistic variable $w^{1/\theta}$, henceforth denoted by x, against t, and estimate by eye the value of A; call this estimate A'. Then plot $\ln [x/(A' - x)]$ against t. If the logistic function is the correct one and A' is also correct, this should give, apart from random errors, a straight line. Now the variance of $\ln [x/(A - x)]$ using the first-order approximation, is given by $[A/(A - x)]^2\sigma^2 = \xi^2\sigma^2$ in the notation already established. This variance increases as x increases, tending to infinity as x tends to A. Thus when the variance of x is constant, the graph of $\ln [x/(A' - x)]$ against t will show increasing scatter as x approaches A. In addition, if x has a symmetrical distribution, that of $\ln [x/(A - x)]$ will be positively skewed for x near A and positive deviations will be in mean value much greater than negative ones. If A' is too high, then the graph of $\ln [x/(A' - x)]$ against t will show a tendency to level off at $\ln [A/(A' - A)]$, while if A' is too low, the graph will show an increasing slope as x approaches A. If the graph shows a sigmoid shape, the logistic equation is inappropriate. Assuming that no sigmoid shape has been found, we may obtain an improved estimate of A as follows. Divide the points into two sets, with the dividing line roughly where $x = A'/2$; then points with $x < A'/2$ are relatively slightly affected by errors in A, and have the smallest variances. Fit by eye a straight line to the points on the $(\ln [x/(A' - x)], t)$-graph using the points where $x < A'/2$ and giving most weight to the smaller values of x. Read off the fitted values of $x/(A - x)$ on this line corresponding to the values of t for points having $x >$ A$'/2$. If z is such a fitted value, then, ignoring errors, $z = x/(A - x)$ and so $A = x(1 + z)/z$. Hence if we write $u = x(1 + z)/z$, then u is an estimate of A. To obtain an improved value for A, calculate u for each point having $x > A'/2$ and use $\bar{u}$, the simple mean of the u's. The method is quite rapid in practice if a logarithmic graph paper is used, since then $\ln [x/(A' - x)]$ can be plotted directly from a knowledge of $x/(A' - x)$ and z can also be read off directly from the fitted line. If the improved value of A still gives curvature, the process may be repeated. Having obtained a good value for A_0, the starting value for A, we may now calculate λ_0 and κ_0 either by a graphical fitting of a straight line to the graph of $\ln [x/(A_0 - x)]$ against t, using the relation $\ln [X/(A_0 - X)] = \lambda + \kappa t$ or by a weighted regression of $\ln [x/(A_0 - x)]$ on t using as weights $\xi_0^2 = [(A_0 - x)/A_0]^2$. When

TABLE IV

Calculations to Obtain Starting Values for Fitting the Logistic Curve

t	x	$x/(71 - x)$	$u = x(1 + z)/z$	$x/(73.2 - x)$
−2.15	3.57	0.0529		0.0513
−1.50	6.25	0.0965		0.0934
−0.85	9.54	0.155		0.150
−0.08	16.91	0.313		0.300
+0.52	24.51	0.510		0.503
1.10	33.78	0.908		0.857
2.28	50.00	2.38	71.9	2.16
3.23	62.05	6.93	74.0	5.57
4.00	69.34	41.8	76.5	18.0
4.65	67.09	17.2	71.1	11.0
5.00	69.34	41.8	72.4	18.0
			Mean 73.2	

σ^2 is small, the fit by eye should be quite adequate, but the weighted regression may have some advantage when the scatter about the line is considerable. If logarithmic paper is used, it must be remembered that one cycle corresponds to $\log_e 10 = 2.30259$ units when the slope is to be measured.

Worked example

The data in Table IV relate to the growth of carrot tops in a field experiment, and were obtained by my colleague Mr. R. B. Austin. θ has been taken as known and equal to 2, so that x, the logistic variable, equals $W^{-1/2}$; t is a time scale based on total incoming radiation (negative values occur because an origin had been taken with a zero between

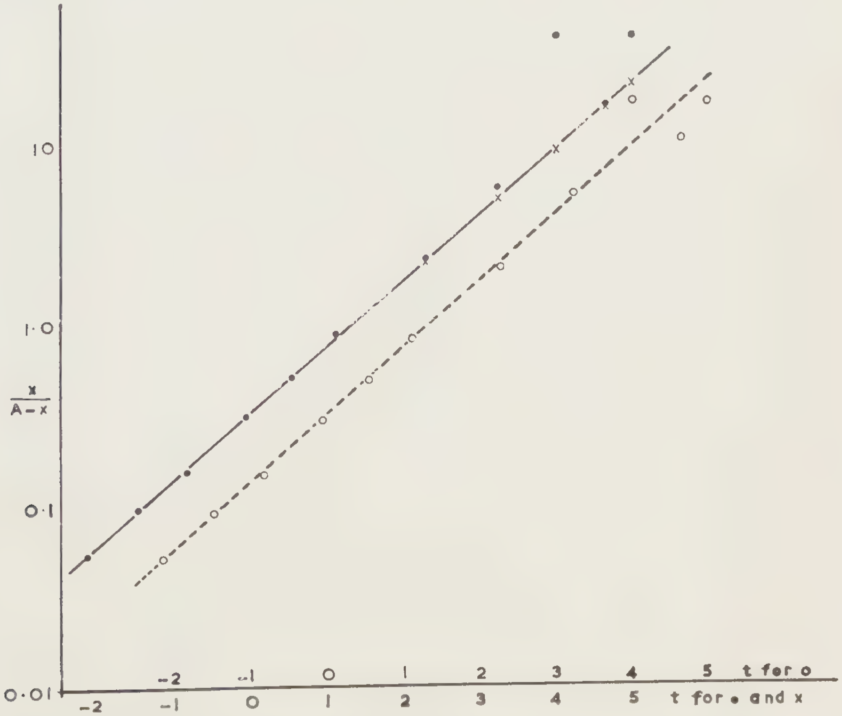


FIGURE 1

METHOD OF OBTAINING STARTING VALUES FOR α, λ, κ, FROM DATA OF TABLE IV.

● Values of $x/(71 - x)$ plotted on logarithmic scale against t.

× Fitted values of $Z = (x/A - x)$ obtained from line drawn by eye through points having $x < 35.5$.

○ Values of $x/(73.2 - x)$ plotted on logarithmic scale against t.
[Note: t— scale displaced to avoid overlapping]

- - - Line fitted by eye through points ○.

the fourth and fifth reading; the reasons for this have no relevance for this example). Table IV also contains the preliminary calculations necessary to arrive at A_0, the starting value for A. An initial plot of x against t gave an A', as judged by eye, of 71. $x/(71 - x)$ was then calculated and recorded in the third column of Table IV. Figure 1 shows the graph of $\ln [x/(71 - x)]$ against t. From the solid line fitted by eye to the points having $x < A'/2 = 35.5$, the fitted values of $x/(A - x)$ for the last 5 points were read off, with u calculated for each, and recorded in the fourth column of Table IV. The revised value of A' is $\bar{u} = 73.2$. Values of $x/(73.2 - x)$ were then calculated (fifth column of Table IV) and plotted against t (Figure 1); the broken line was fitted by eye, giving most weight to the small values of x, and gave $\lambda_0 = -1.109$, $\kappa_0 = 0.861$. The calculations for the exact iterative method discussed above can now be started. Filling up Table III we have the following:

t_i	τ_i	ξ_i	$\xi_i t_i$	$\ln (1 - \xi_i)$	y_i	$y_i - Y_i$
−2.15	−2.960	0.9507	−2.0440	−3.0105	1.272	−0.0020
−1.50	−2.400	0.9168	−1.3752	−2.4868	1.832	+0.0343
−0.85	−1.841	0.8631	−0.7336	−1.9883	2.255	−0.0412
−0.08	−1.178	0.7646	−0.0612	−1.4464	2.828	−0.0101
+0.52	−0.661	0.6595	+0.3429	−1.0773	3.199	−0.0082
+1.10	−0.162	0.5404	0.5944	−0.7774	3.520	+0.0129
+2.28	+0.854	0.2986	0.6808	−0.3547	3.912	−0.0178
+3.23	+1.672	0.1582	0.5110	−0.1722	4.128	+0.0157
+4.00	+2.335	0.0883	0.3532	−0.0924	4.239	+0.0469
+4.65	+2.895	0.0524	0.2437	−0.0538	4.206	−0.0247
+5.00	+3.196	0.0393	0.1965	−0.0401	4.239	−0.0054

$\bar{y} = 3.2391, \sum(1 - \xi_i)/n = -1.0454, \alpha_0 = 4.2845, S.S. = 0.00661578.$

Adjustment equations are

$$\begin{bmatrix} 11.0000 & 5.3319 & -1.2915 \\ & 3.9272 & -3.0008 \\ & & 8.0293 \end{bmatrix} \begin{bmatrix} \delta\alpha_0 \\ \delta\lambda_0 \\ \delta\kappa_0 \end{bmatrix} = \begin{bmatrix} 0.0004 \\ -0.0124 \\ -0.0020 \end{bmatrix}.$$

Solution is given by

$$\begin{array}{rlcrl} \delta\alpha_0 = & 0.010025, & & \alpha_1 = & 4.2940, \\ \delta\lambda_0 = & -0.021566, & \text{from whence} & \lambda_1 = & -1.1300, \\ \delta\kappa_0 = & -0.006746, & & \kappa_1 = & 0.8544. \end{array}$$

If a further iteration is required then α_1 should be estimated in the same way as α_0, i.e. using $\bar{y} - \sum \ln (1 - \xi_i)/n$, and the value obtained above ignored. However the adjustments obtained from the first cycle are sufficiently small to justify our stopping in this case.

The goodness of fit of the logistic curve can now be checked by comparing the residual mean square after fitting with the residual error obtained from replication. Using the estimates α_1, λ_1, κ_1, we have a residual sum of squares after fitting of 0.00642 giving a mean square with $11 - 3 = 8$ d.f. of 0.00080. The error mean square was 0.0033 so that the fit is satisfactory. We may now obtain the dispersion matrix by inverting the information matrix and multiplying it by the variance of a single y_i, as estimated by the error mean square. Thus the dispersion matrix is given by

$$0.0033 \times \begin{bmatrix} 11.0000 & 5.3319 & -1.2915 \\ & 3.9272 & -3.0008 \\ & & 8.0293 \end{bmatrix}^{-1} = \begin{bmatrix} 0.00137 & -0.00237 & 0.00066 \\ & 0.00526 & 0.00159 \\ & & 0.00090 \end{bmatrix}.$$

The standard errors of the estimates affect the second place of decimals, thus justifying the stopping of the iterative process after one cycle.

FITTING THE GENERALIZED LOGISTIC CURVE

If we now define τ_i as $(\lambda + \kappa t_i)/\theta$, then from (2)

$$Y = \alpha + \theta \ln (1 - \xi)$$

where $\xi = e^{-\tau}/(1 + e^{-\tau})$ as before, and

$$\frac{\partial Y_i}{\partial \alpha} = 1, \quad \frac{\partial Y_i}{\partial \lambda} = \xi_i, \quad \frac{\partial Y_i}{\partial \kappa} = \xi_i t_i, \quad \frac{\partial Y_i}{\partial \theta} = \ln (1 - \xi_i) - \tau_i \xi_i = \beta_i, \text{ say.}$$

The least-square equations, therefore, are given by

$$\begin{aligned} &\sum (y_i - Y_i) = 0, \qquad \sum (y_i - Y_i)\xi_i t_i = 0, \\ &\sum (y_i - Y_i)\xi_i = 0, \qquad \sum (y_i - Y_i)\beta_i = 0 \end{aligned} \tag{6}$$

and the iterative solution from starting values α_0, λ_0, κ_0, θ_0, gives adjustments $\delta\alpha_0$ etc. from the solution of

$$\begin{bmatrix} n & \sum \xi_i & \sum \xi_{iti} & \sum \beta_i \\ & \sum \xi_i^2 & \sum \xi_i^2 t_i & \sum \beta_i \xi_i \\ & & \sum \xi_i^2 t_i^2 & \sum \beta_i \xi_i t_i \\ & & & \sum \beta_i^2 \end{bmatrix} \begin{bmatrix} \delta\alpha_0 \\ \delta\lambda_0 \\ \delta\kappa_0 \\ \delta\theta_0 \end{bmatrix} = \begin{bmatrix} \sum (y_i - Y_i) \\ \sum (y_i - Y_i)\xi_i \\ \sum (y_i - Y_i)\xi_i t_i \\ \sum (y_i - Y_i)\beta_i \end{bmatrix}. \tag{7}$$

TABLE V

VALUES OF $-10^4\beta$, $-\beta = \tau/(1 + e^{-\tau}) - \ln(1 + e^{\tau})$

τ	0	1	2	3	4	5	6	7	8	9
0.0	6931	6931	6931	6930	6929	6928	6927	6925	6923	6921
0.1	6919	6916	6913	6910	6907	6903	6900	6895	6891	6887
0.2	6882	6877	6871	6866	6860	6854	6848	6841	6834	6827
0.3	6820	6813	6805	6797	6789	6781	6772	6763	6754	6745
0.4	6735	6726	6716	6706	6695	6685	6674	6663	6652	6640
0.5	6628	6617	6605	6592	6580	6567	6554	6541	6528	6515
0.6	6501	6487	6473	6459	6445	6430	6415	6400	6385	6370
0.7	6355	6339	6323	6307	6291	6275	6258	6242	6225	6208
0.8	6191	6174	6157	6139	6122	6104	6086	6068	6050	6031
0.9	6013	5994	5976	5957	5938	5919	5900	5881	5861	5842
1.0	5822	5802	5782	5763	5743	5722	5702	5682	5662	5641
1.1	5620	5600	5579	5558	5537	5516	5495	5474	5453	5432
1.2	5411	5389	5368	5346	5325	5303	5281	5260	5238	5216
1.3	5194	5172	5150	5128	5106	5084	5062	5040	5018	4996
1.4	4974	4951	4929	4907	4885	4862	4840	4818	4795	4773
1.5	4751	4728	4706	4683	4661	4639	4616	4594	4571	4549
1.6	4527	4504	4482	4460	4437	4415	4393	4370	4348	4326
1.7	4304	4282	4259	4237	4215	4193	4171	4149	4127	4105
1.8	4083	4061	4039	4018	3996	3974	3952	3931	3909	3887
1.9	3866	3844	3823	3802	3780	3759	3738	3717	3695	3674
2.0	3653	3632	3611	3591	3570	3549	3528	3508	3487	3467
2.1	3446	3426	3406	3385	3365	3345	3325	3305	3285	3265
2.2	3245	3226	3206	3186	3167	3147	3128	3109	3090	3070
2.3	3051	3032	3013	2994	2976	2957	2938	2920	2901	2883
2.4	2864	2846	2828	2810	2792	2774	2756	2738	2721	2703
2.5	2685	2668	2650	2633	2616	2599	2582	2565	2548	2531
2.6	2514	2497	2481	2464	2448	2431	2415	2399	2383	2367
2.7	2351	2335	2319	2303	2288	2272	2257	2241	2226	2211
2.8	2195	2180	2165	2150	2136	2121	2106	2091	2077	2062
2.9	2048	2034	2020	2005	1991	1977	1964	1950	1936	1922
3.0	1909	1895	1882	1868	1855	1842	1829	1816	1803	1790
3.1	1777	1764	1752	1739	1726	1714	1702	1689	1677	1665
3.2	1653	1641	1629	1617	1605	1594	1582	1570	1559	1547
3.3	1536	1525	1514	1502	1491	1480	1469	1459	1448	1437
3.4	1426	1416	1405	1395	1384	1374	1364	1354	1344	1333
3.5	1323	1314	1304	1294	1284	1274	1265	1255	1246	1236
3.6	1227	1218	1208	1199	1190	1181	1172	1163	1154	1146
3.7	1137	1128	1120	1111	1102	1094	1086	1077	1069	1061
3.8	1053	1045	1037	1029	1021	1013	1005	997	989	982
3.9	974	967	959	952	944	937	930	922	915	908
4.0	901	894	887	880	873	866	859	853	846	839
4.1	833	826	820	813	807	800	794	788	782	775
4.2	769	763	757	751	745	739	733	728	722	716

TABLE V—(*Continued*)

τ	0	1	2	3	4	5	6	7	8	9
4.3	710	705	699	694	688	682	677	672	666	661
4.4	656	650	645	640	635	630	625	620	615	610
4.5	605	600	595	590	586	581	576	572	567	562
4.6	558	553	549	544	540	536	531	527	523	518
4.7	514	510	506	502	498	494	490	486	482	478
4.8	474	470	466	462	458	455	451	447	444	440
4.9	436	433	429	426	422	419	415	412	408	405
5.0	402	398	395	392	389	386	382	379	376	373
5.1	370	367	364	361	358	355	352	349	346	343
5.2	340	337	335	332	329	326	324	321	318	316
5.3	313	310	308	305	303	300	298	295	293	290
5.4	288	285	283	281	278	276	274	271	269	267
5.5	265	262	260	258	256	254	252	249	247	245
5.6	243	241	239	237	235	233	231	229	227	225
5.7	223	222	220	218	216	214	212	211	209	207
5.8	205	204	202	200	198	197	195	193	192	190
5.9	188	187	185	184	182	181	179	178	176	175
6.0	173	172	170	169	167	166	164	163	162	160
6.1	159	158	156	155	154	152	151	150	148	147
6.2	146	145	143	142	141	140	138	137	136	135
6.3	134	133	132	130	129	128	127	126	125	124
6.4	123	122	121	120	119	118	116	116	114	114
6.5	113	112	111	110	109	108	107	106	105	104
6.6	103	102	101	101	100	99	98	97	96	95
6.7	95	94	93	92	91	91	90	89	88	88
6.8	87	86	85	85	84	83	82	82	81	80
6.9	80	79	78	77	77	76	75	75	74	73
7.0	73	72	72	71	70	70	69	69	68	67

Little amendment is needed to Table III for the solution of these equations. A column for β must be inserted between columns 4 and 5, and β can be obtained rapidly from Table V. α_0 is estimated as $\bar{y} - \theta_0 \cdot \sum \ln (1 - \xi_i)/n$, and $y_i - Y_i = y_i - \alpha_0 - \theta_0 \ln (1 - \xi_i)$. All the elements in (7) can then be calculated.

Obtaining the starting values

A sign that θ is not equal to unity is the appearance of a trend in the values of u during the graphical process for finding A_0 . If $\theta > 1$, then values of u show a decreasing trend as w departs from A, and the lowest of them are often below the highest values of w. Conversely if $\theta < 1$, then u shows an increasing trend as w falls. According to the trend shown by the u's, θ can be changed in the appropriate direction, and when θ_0 has been obtained, with the corresponding $A^{1/\theta}$, we can use the relation

$$E\left[\ln \frac{x^{1/\theta}}{A^{1/\theta} - x^{1/\theta}}\right] \sim \frac{1}{\theta}(\lambda + \kappa t) = \tau$$

to get estimate of λ_0/θ_0 and κ_0/θ_0 and hence of λ_0 and κ_0 . For the calculations of coefficients in (7), however, only t and τ are needed.

Worked example

In the example above of fitting the logistic curve to $W^{1/2}$, we were in fact fitting the generalized logistic curve taking θ as known and equal to 2. We now consider the fitting when θ is not assumed known, using the value 2 as a starting point. Columns 1–5 of Table III are unchanged but y in column 6 is now twice the previous value since we were working before with $\ln x^{1/2} = \frac{1}{2} \ln x$. Column 7 will also be doubled and a column of β_i must be added. The revised and added columns are given together with the iterative equations in Table VI.

TABLE VI

REVISED AND ADDITIONAL COLUMNS FOR FITTING THE GENERALIZED LOGISTIC CURVE

β_i	y_i	$y_i - Y_i$
−0.1964	2.544	−0.0038
−0.2865	3.664	+0.0688
−0.3992	4.510	−0.0824
−0.5457	5.656	−0.0200
−0.6414	6.398	−0.0162
−0.6895	7.040	+0.0252
−0.6096	7.824	+0.0356
−0.4366	8.256	+0.0314
−0.2987	8.478	+0.0942
−0.2054	8.412	−0.0494
−0.1657	8.478	−0.0106

Adjustment equations are

$$\begin{bmatrix} 11.0000 & 5.3319 & -1.2915 & -4.4747 \\ & 3.9273 & -3.0008 & -2.3015 \\ & & 8.0293 & -0.3343 \\ & & & 2.1857 \end{bmatrix} \begin{bmatrix} \delta\alpha_0 \\ \delta\lambda_0 \\ \delta\kappa_0 \\ \delta\theta_0 \end{bmatrix} = \begin{bmatrix} 0.0016 \\ -0.0244 \\ -0.0048 \\ 0.0096 \end{bmatrix}.$$

The solution of these equations is given by

$$\begin{aligned} \delta\alpha_0 &= -0.016240, \\ \delta\lambda_0 &= -0.201431, \\ \delta\kappa_0 &= -0.089091, \\ \delta\theta_0 &= -0.254585, \end{aligned}$$

whence

$$\begin{aligned} \lambda_1 &= \theta_0\left(\frac{\lambda_0}{\theta_0}\right) + \delta\lambda_0 = 2(-1.109) - 0.2014 = -2.4194, \\ \kappa_1 &= \theta_0\left(\frac{\kappa_0}{\theta_0}\right) + \delta\kappa_1 = 2(0.861) - 0.0891 = +1.6329, \\ \theta_1 &= 2 - 0.2546 = 1.7454. \end{aligned}$$

α_1 is not given because it is better estimated from $\bar{y} - \theta_1 \sum \ln(1 - \xi_i)/n$ in the next iteration, as described above.

The rather large and approximately equal adjustments to λ_0 and θ_0 should be noted; the estimates of these parameters are in fact highly correlated and have relatively large variances; hence the reduction of the *S. S.* of residuals (from 0.0265 to 0.0239) following the adjustment is rather small.

A second iteration using the values α_1 , λ_1 , κ_1 , θ_1 gives

$$\begin{aligned} \delta\alpha_1 &= -0.00446770, & \alpha_2 &= 8.5538, \\ \delta\lambda_1 &= +0.02565976, \quad \text{and} & \lambda_2 &= 2.3937, \\ \delta\kappa_1 &= +0.00782844, & \kappa_2 &= 1.6407, \\ \delta\theta_1 &= +0.02010622, & \theta_2 &= 1.7655. \end{aligned}$$

and a third and final iteration gives

$$\begin{aligned} \alpha_3 &= 8.5539, \\ \lambda_3 &= 2.3917, \\ \kappa_3 &= 1.6415, \\ \theta_3 &= 1.7676 \end{aligned}$$

with dispersion matrix

$$\begin{bmatrix} 0.01419 & 0.02379 & 0.01526 & 0.06075 \\ & 0.13220 & 0.06751 & 0.21618 \\ & & 0.03705 & 0.11775 \\ & & & 0.40847 \end{bmatrix}.$$

In comparing the estimates and their variances and covariances with those obtained for θ assumed known and equal to 2, it should be remembered that the effective parameters estimated in the latter case were $\alpha/\vartheta = \alpha/2$ etc. Thus parameter values for the logistic case should be multiplied by 2 and their variances and covariances by 4 for comparison with those of the four-parameter curve. As was to be expected, the variances of $\hat{\alpha}$, $\hat{\lambda}$, and $\hat{\kappa}$ have been considerably increased by the inclusion of θ, also $\hat{\theta}$ itself is poorly determined with a standard error of ± 0.639, so that it does not differ significantly from the value of 2 originally used.

ALLOCATION OF SAMPLE POINTS

This section is concerned with the effect on the efficiency of estimation of different allocations of position of sample points on the curve. Obviously points on the curve where $x^{1/\theta}$ is small compared to $A^{1/\theta}$ give almost no information about A, while points where $x^{1/\theta}$ is close to $A^{1/\theta}$ give almost no information about λ and κ, the parameters defining the exponential part of growth. To reduce the range of possibilities of the distribution of the sample points to manageable proportions, attention will be confined to samples at points equally spaced with regard to t and symmetrically arranged about $\tau = 0$. [In practice, a rough idea of the values of the parameters will enable observations to be taken at times similar to those considered here]. The points of the arrangements considered are spaced along the t scale in units of τ or $\frac{1}{2}\tau$, while the ranges spanned are 6τ, 10τ, and 14τ. Table VII gives the relative sampling variances per point for the parameters, and also the generalized information per point using $|\mathbf{Q}|/n$ where n is the number of points and $\mathbf{Q}$ the information matrix for the samples. Two cases are considered: θ known (3 parameters to be estimated) and θ unknown (4 parameters to be estimated).

The table shows that more intensive sampling of a given range [e.g. $-3(\frac{1}{2})3$ instead of $-3(1)3$] slightly increases the sampling variance per point while increasing the information per point. The increased information implies that some or all of the correlations between the estimates have been reduced by the increased density of points, although the arrangement $3(\frac{1}{2})3$ is less efficient than $3(1)3$ when measured in terms of sampling variances only. Increasing the range reduces the sampling variances per point as one might expect, and so yields more information than more intensive sampling of the same range. The sampling variances of $\hat{\alpha}$, $\hat{\lambda}$, and $\hat{\kappa}$ are considerably increased when θ is unknown, particularly when the range of τ is small. This is of course a reflection of the fact that curves for neighbouring values of

TABLE VII

Sampling Variances Per Point of the Parameter Estimates and the Generalized Information Per Point for Various Distributions of the Observations

θ Known, α, λ, κ Unknown, $\sigma^2 = 1$

No. of points	Allocation of τ	Sampling variance per point			Generalized information per point
		$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\kappa}$	
7	$-3(1)3$	5.39	39.2	3.60	1.57
11	$-5(1)5$	3.11	23.1	1.14	16.93
15	$-7(1)7$	2.59	18.5	0.56	73.16
13	$-3(\frac{1}{2})3$	6.05	42.9	4.21	4.32
21	$-5(\frac{1}{2})5$	3.22	23.8	1.25	54.80
29	$-7(\frac{1}{2})7$	2.63	18.8	0.60	252.6

α, λ, κ, θ Unknown, $\sigma^2 = 1$

No. of points	Allocation of τ	Sampling variance per point				Generalized information per point
		$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\kappa}$	$\hat{\theta}$	
7	$-3(1)3$	14.26	361.2	41.50	339.3	0.0324
11	$-5(1)5$	4.99	84.0	5.28	89.4	2.008
15	$-7(1)7$	3.50	53.4	1.62	54.4	20.16
13	$-3(\frac{1}{2})3$	17.20	420.8	51.57	398.4	0.2655
21	$-5(\frac{1}{2})5$	5.34	91.5	6.13	96.3	11.95
29	$-7(\frac{1}{2})7$	3.60	45.5	1.82	56.2	130.4

θ can be made to agree very closely over this range of τ by suitable changes in the other parameters. $\hat{\lambda}$ and $\hat{\theta}$ have the largest sampling variances in all the arrangements and hence are the most difficult to estimate accurately. For instance, if an individual value of w_i in an experiment had a percentage standard error of 5, then var y = var (ln w) ~ 0.0025, so that with the allocation $-3(1)3$ we should have var $\hat{\theta} = 0.0025 \times 48.47$, whence s.e. $\hat{\theta} = \pm 0.35$. Since θ is often in the range 1–2, the allocation $\tau = -3(1)3$ would be inadequate to determine this parameter at all accurately. Similar considerations apply to $\hat{\lambda}$. By such arguments one can find what range w must cover if adequate estimates of the parameters are to be obtained for

$$W = A/(1 + e^{-\tau})^{\theta}$$

so that $\tau = \ln [W^{1/\theta}/(A^{1/\theta} - W^{1/\theta})] = \text{logit } (W/A)^{1/\theta}$. Thus if θ is, say, equal to 2 then $\tau = 3$ corresponds to $(W/A)^{1/\theta} = 0.95257$ whence $W/A = 0.976$; similarly $\tau = -3$ gives $W/A = 0.024$. Thus when $\theta = 2$ the range $\tau = -3$ to 3 implies that W ranges from 2.4 to 97.6 percent of A, and the argument above shows that this range will often be insufficient for adequate estimates of some of the parameters. Note that the larger the value of θ the greater the range of W/A required to cover the same range of τ.

DISCUSSION

This paper is not the place for a discussion of the place of mathematical functions in the description and analysis of growth. A great deal has been written on the subject both for and against the utility of this approach, and it suffices to say that I assume in this paper that differential equations and their associated integrals *have* some usefulness, and that methods are required for the efficient estimation of the parameters. For an approach to the problem of interpreting growth curves which involves replacing chronological time t by some function of meteorological observations see Nelder *et al.* [1960]. [In the example above t is in fact a scale based on total incoming radiation].

The method of estimation used in this paper is based on three assumptions: (i) that the w_i are independent, (ii) that $E(y_i) = Y_i$ and (iii) that var y_i is independent of Y_i. It may reasonably be asked how suitable the procedure is likely to be if some or all of the assumptions are not true, and this involves us in an analysis of the errors in the Y_i. Errors may be due to variation in the parameters from one individual to another, for either genetic or non-genetic reasons; or they may arise from errors in t_i, caused by using the wrong time-scale. In addition we have the possibility of systematic errors arising from deviations from the form of the curve actually used in the fitting. It will certainly not be exactly true that if, for instance, κ has a normal distribution in the population of individuals whose growth is being studied, then random sampling will enable us to set $E(y_i) = Y_i$, where Y_i is given by (3) with the mean of κ used in the formula. But such biases will not be serious if the variance of κ is small, i.e. if we control the genetic and environmental uniformity of the population well. If the w_i are from the same experimental units throughout, which is the commonest cause of non-independence, the method will remain satisfactory if var (y_i) remains constant, where the variance is now measured *within* individual growth curves. However sampling variances of the estimates of parameters must now be calculated from the variation in the parameter values *between* individuals. If the variation between individual curves is much greater than the variation within them, then the efficiency of the fitting method becomes correspondingly less important, and even graphical methods might suffice. A frequent cause of deviations of var (y_i) from constancy is the occurrence of constant-variance weighing errors on the w scale. On the log. scale these become more serious as the weight falls, and lead to var (y_i) increasing rapidly at very low values of w, where the weights are not large compared to the weighing errors. Difficulties in measuring very small distances on a plant cause the same effect. In practice the change in var (y) can often be approximately eliminated by a suitable increase in replication at very low levels of w.

The model employed here for fitting the logistic curve may be contrasted with a method suggested by Stevens [1951], who proposed writing it in the form

$$\frac{1}{W} = \frac{1}{A}(1 + \beta e^{-\kappa t})$$

and using his method for fitting curves of the type

$$z = \alpha + \beta\rho^t, \quad \text{with} \quad z = \frac{1}{W}, \quad \text{and} \quad \rho = e^{-\kappa}.$$

Since he gave equal weight to his z's in the fitting, this is equivalent to assuming that $1/w$ has constant variance and hence that var $(\ln w) \propto W^2$ to the first order. Though conditions may exist where var $(\ln w) \propto W^2$, in my experience it is usually much closer to the truth to take var $(\ln w)$ constant. Several papers concerned with fitting the curve $z = \alpha + \beta\rho^t$ have appeared (e.g. Nair [1954], Patterson [1956], Finney [1958], Patterson and Lipton [1959]), but these are almost entirely concerned with the case where z has constant variance and where the t_i are equally spaced. Where time scales based on meteorological measurements (such as day-degrees) are used for field crops, it is usually impossible or impracticable to arrange for equally spaced t's, so that methods used in the above mentioned papers are no longer available, even if it could be assumed that var z was constant. In addition, none of the methods is immediately applicable to the general case where θ has to be estimated.

The reasons behind the particular choice of the parameters in (3) should perhaps be mentioned. There are, of course, an infinite number of equivalent ways of writing (3), in so far as the structural parameters are concerned. The particular form used has been chosen to reduce the computing labour of the iterative solution as much as possible, while retaining a set of parameters with fairly clear meanings as far as growth is concerned. Thus α = ln (asymptotic value of W), while λ and κ define the exponential growth stage for large negative t when $W \sim Ae^{\lambda+\kappa t}$. A possible interpretation of θ for a field crop in terms of the spatial distribution of plant material has been given by Nelder *et al.* [1960]. The form used may be capable of further improvement, especially from the point of view of the convergence of the iterative process. It is an interesting point to what extent approximately orthogonal parameters can be found for the general non-linear function, that is parameters such that the cross terms in the dispersion matrix are small compared to the diagonal terms, and whether the use of such parameters would speed the convergence of the iterative process. While a linear transformation of the parameters leads to equivalent iterative equations, non-linear transformations in general do not, and possibly the speed of convergence would be affected by the particular specification of the parameters.

A natural extension of the generalized logistic equation as defined here is the five-parameter equation $dW/dt = \alpha W^{\xi} - \beta W^{\eta}$ discussed by von Bertalanffy [1957]. The estimation of the parameters of this equation, subject to the same assumptions as those considered above, is greatly complicated by the lack of any general explicit solution of the equation. Consequently expected values would have to be computed by numerical integration which makes the labour involved excessive except to those having electronic computers available. A general method for the fitting of differential equations has been given by Box [1956], and this could be applied to the general equation. However we may expect that the introduction of the fifth parameter would entail the existence of very extensive data if reasonably accurate estimates of the parameters were to be obtained. The data discussed above cover a 380-fold range of weights, and the standard error of a single weight was about 5.5 percent, yet the s.e. of $\hat{\theta}$ was as high as ±0.64. The introduction of the fifth parameter might well require a much larger range of weights to give reasonably accurate estimates. One consequence of this situation is that it becomes difficult to test models based on *a priori* values of the exponents ξ and η, unless very extensive data are available, since the fits of curves having ξ and η differing quite widely from the *a priori* values would be hardly distinguishable.

SUMMARY

The least-squares fit of the curves defined by

$$\frac{dW}{dt} = \kappa W\left[1 - \left(\frac{W}{A}\right)^{1/\theta}\right]$$

is derived for the case when the sample values of ln W are independent, unbiased, and of constant variance.

Tables are provided to assist the computing of the iterative process used for estimation, and a method is given for obtaining starting values for the parameters.

Evaluation of the sampling variances of the estimates of the parameters and the generalized information shows that the range which the observations cover is more important than the density of sample points within that range if minimum sampling variance or maximum information per point is required. The introduction of θ as an unknown increases the sampling variances of the other parameters, the increase being especially marked when the range of $(W/A)^{1/\theta}$ is small.

Some possible sources and consequences of deviations from the assumptions underlying the fitting are discussed, and also the difficulties involved in extending the fitting to the curves defined by

$$dW/dt = \alpha W^{\xi} - \beta W^{\eta}.$$

It is stressed that sensitive tests of hypotheses involving *a priori* values of parameters may demand very extensive data.

ACKNOWLEDGEMENT

I should like to thank the referees for several useful suggestions concerning the presentation of this paper.

REFERENCES

Bailey, N. T. J. [1951]. Testing the solubility of maximum likelihood equations in the routine application of scoring methods. *Biometrics 7*, 268–74.

Berkson, J. [1953]. A statistically precise and relatively simple method of estimating the bio-assay with quantal response based on the logistic function. *J. Amer. Stat. Assoc. 48*, 565–99.

Bertalanffy, L. von [1957]. Quantitative laws in metabolism and growth. *Quart. Rev. Biol. 32*, 218–31.

Box, G. E. P. [1956]. *Some notes on non-linear estimation.* Mimeographed notes, Princeton University.

Comrie, L. J. [1949]. *Chambers' six-figure mathematical tables, Vol. II.* Chambers, Edinburgh.

Cornfield, J. and Mantel, N. [1950]. Some new aspects of the application of maximum likelihood to the calculation of the dosage response curve. *J. Amer. Stat. Assoc. 45*, 181–210.

Dwyer, P. S. [1951]. *Linear Computations.* Wiley, New York.

Finney, D. J. [1958]. The efficiencies of alternative estimates for an asymptotic regression equation. *Biometrika 45*, 370–88.

Hartley, H. O. [1948]. The estimation of non-linear parameters by internal least-squares. *Biometrika 35*, 32–45.

Nair, K. R. [1954]. The fitting of growth curves. *Statistics and Mathematics in Biology* 119–32. Iowa State College Press.

Nelder, J. A., Austin, R. B., Bleasdale, J. K. A., and Salter, P. J. [1960]. An approach to the study of yearly and other variation in crop yields. *J. Hort. Sci. 35*, 73–82.

Patterson, H. D. [1956]. The use of autoregression in fitting an exponential curve. *Biometrika 45*, 389–400.

Patterson, H. D., and Lipton, S. [1959]. An investigation of Hartley's method for fitting an exponential curve. *Biometrika 46*, 281–92.

Pütter, A. [1920]. Studien über physiologische Ähnlichkeit. VI, Wachstumsähnlichkeiten. *Arch. ges. Physiol. 180*, 298–340.

Richards, F. J. [1959]. A flexible growth function for empirical use. *J. Exper. Bot. 10* 290–300.

Stevens, W. L. [1951]. Asymptotic regression. *Biometrics 7*, 247–67.

COMBINED ANALYSIS OF BALANCED INCOMPLETE BLOCK DESIGNS WITH SOME COMMON TREATMENTS

M. V. Pavate
Central Tobacco Research Institute
Rajahmundry, India

INTRODUCTION

On many occasions, research workers have to lay out their experiments at different places with one or more treatments common to each experiment. The reason may be the lack of sufficient space at one location as in field experiments, or as in industrial problems the availability of only a limited number of experimental units with stipulated conditions with the result that the experiment has to be laid out at different plants. In either case, the research workers are interested in the combined analysis of the data from all of the experiments, besides their individual analyses. The purpose of the present paper is to suggest a simplified method to obtain adjusted treatment components for the combined analysis when the individual experiments are laid out in balanced incomplete block (BIB) designs. Gomes and Guimares [1958] have considered the case when the individual experiments are laid out in randomised complete block designs. The case dealt with by Gomes and Guimares could be deduced as a special case from general method suggested in this paper.

NATURE OF THE PROBLEM

Suppose there are g BIB experiments to be analysed jointly with the following parameters:

v = number of treatments,
b = number of blocks,
r = number of replications for each treatment,
k = number of units per block,
λ = number of times any two treatments occur together in the design.

Let us suppose that, for each experiment, there are c treatments common. Hence, there are $c + g(v - c)$ different treatments in all. Let us, following Gomes and Guimares, call the c treatments (common to all experiments) 'common' treatments and the rest, $g(v - c)$ treatments, 'regular' treatments. Hence, the combined experiment will have for its parameters

$$\begin{aligned} v' &= c + g(v - c), \qquad b' = gb, \qquad k' = k, \\ r' &= gr \text{ for the common treatments,} \\ &\quad\; r \text{ for the regular treatments,} \\ \lambda' &= g\lambda \text{ for two common treatments,} \\ &\quad\; \lambda \text{ for a common and a regular treatment from the same experiment,} \\ &\quad\; 0 \text{ for two regular treatments from two different experiments.} \end{aligned}$$

As one can see, the design is no longer a balanced one, but still possesses the property of connectedness.

METHODS OF OBTAINING ADJUSTED TREATMENT SUM OF SQUARES

We proceed to find the adjusted treatment sum of squares for the combined analysis by two methods. The general method suggested by C. R. Rao [1947] treats the combined data as that of a new incomplete block experiment. Here we will only outline this method since it is well known in the statistical literature. The simplified method, which the author would like to introduce, makes use of the individual analyses already performed and this will be given later, along with the results for various special cases.

Classical Method

Here, the whole problem of combined analysis reduces to the solution of the adjusted normal equations

$$\mathbf{Ct} = \mathbf{Q} \tag{1}$$

where the matrix $\mathbf{C}$ $(v' \times v')$ and the column vectors $\mathbf{t}$ $(v' \times 1)$, $\mathbf{Q}$ $(v' \times 1)$ have usual meanings. The elements of the $\mathbf{C}$ matrix are

$$\begin{aligned} c_{ii} &= g\lambda(v - 1)/k && \text{for common treatments,} \\ c_{ij} &= -\lambda g/k && \text{for any two common treatments,} \\ c_{ii} &= \lambda(v - 1)/k && \text{for regular treatments,} \\ c_{ij} &= -\lambda/k && \text{for any two regular treatments from the same experiment,} \\ c_{ij} &= 0 && \text{for any two regular treatments from different experiments.} \end{aligned}$$

Here the rank of the $\mathbf{C}$ matrix is $v' - 1$ and hence we adopt a suitable restriction, so that it becomes non-singular and easy to invert. We use the restriction

$$\frac{\lambda g}{k}(t_1 + t_2 + \cdots + t_c) + \frac{\lambda}{k}\sum_{i=c+1}^{g(v-c)} t_i = 0.$$

By using this restriction, the **C** matrix reduces to

$$\begin{bmatrix} Erg\,\mathbf{I}_1 & \mathbf{0} \\ -\frac{\lambda}{k}\mathbf{J}_1 & \mathbf{I}_3 \times \left(Er\,\mathbf{I}_2 - \frac{\lambda}{k}\mathbf{J}_2\right) \end{bmatrix}$$

where **I** is the identity matrix, **0** is a matrix having only zero elements, **J** is a matrix having only unity elements, **×** indicates the direct product and E is the efficiency factor of the BIB design; the dimensions of the matrices are

$$\begin{aligned} &I_1 = ec \times c, \qquad I_2 = (v - c) \times (v - c), \\ &I_3 = g \times g, \qquad J_1 = g(v - c) \times c, \qquad\qquad (2) \\ &J_2 = (v - c) \times (v - c), \qquad 0 = c \times g(v - c). \end{aligned}$$

This matrix has a pattern discussed by Roy in one of his papers (3) and hence the inverse can be easily found. The inverse turns out to be

$$\begin{bmatrix} \frac{1}{Erg}\mathbf{I}_1 & \mathbf{0} \\ \frac{1}{gErc}\mathbf{J}_1 & \mathbf{I}_3 \times \left(\frac{1}{Er}\mathbf{I}_2 + \frac{1}{Erc}\mathbf{J}_2\right) \end{bmatrix}$$

where $\mathbf{I}_1$, $\mathbf{I}_2$, $\mathbf{I}_3$, $\mathbf{J}_1$, $\mathbf{J}_2$, $\mathbf{0}$ are of the same dimensions as in (2). Hence, the solution of (1) turns out to be

$$\hat{\mathbf{t}} = \mathbf{C}^{-1}\mathbf{Q}. \qquad (3)$$

Finally, the adjusted treatment sum of squares is $\hat{\mathbf{t}}'\mathbf{Q}$ where $\hat{\mathbf{t}}'$ $(1 \times v')$, a row vector. Total and Blocks (unadj) sums of squares are found in usual manner.

Simplified method:

This method is useful to those research workers who are not conversant with matrix algebra and advanced mathematical topics. The method suggested here is based on individual analyses of g BIB experiments.

Suppose we have performed individual analyses of g BIB experiments with the same parameter values, viz. v, b, r, k, λ and with c common treatments.

Let

$Q_{i,s}$ = adjusted treatment total of the ith common treatment in sth experiment,
$i = 1, \cdots, c,$
$s = 1, \cdots, g,$
Q_i = $\sum_{s=1}^{g} Q_{i,s}$ = adjusted treatment total for ith common treatment in combined analysis, $i = 1, \cdots, c,$
$Q_j^{(s)}$ = adjusted treatment total for jth regular treatment in sth experiment, $j = 1, \cdots, (v - c)$, $s = 1, \cdots, g.$

From these values, we obtain adjusted treatment sum of squares as follows.

Form the following two-way table, Table I.

TABLE I
Two-Way Table of Common Treatments and Experiments with $Q_{i,s}$ Values

Expts.	c. treat				Total
	t_1	t_2	$\cdots$	t_c	
E_1	$Q_{1,1}$	$Q_{2,1}$	$\cdots$	$Q_{c,1}$	$Q_{.,1}$
E_2	$Q_{1,2}$	$Q_{2,2}$	$\cdots$	$Q_{c,2}$	$Q_{.,2}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
E_g	$Q_{1,g}$	$Q_{2,g}$	$\cdots$	$Q_{c,g}$	$Q_{.,g}$
Total	Q_1	Q_2	$\cdots$	Q_c	$Q_{..}$

Here

$$Q_{.s} = \sum_{i=1}^{c} Q_{i,s} ,$$

$$Q_i = \sum_{s=1}^{g} Q_{i,s} ,$$

and

$$Q_{..} = \sum_{i=1}^{c} \sum_{s=1}^{g} Q_{i,s} .$$

Obtain (common treatments × experiments) interaction in the usual way. Then, the adjusted treatment sum of squares for the combined analysis is Adj. Treat. S.S. (Combined Analysis) = Sum of Adj. Treat. S.S. from each exp.

$$-1/Er \text{ (Common Treat.} \times \text{Exp. Interaction)} \tag{4}$$

ESTIMATION OF TREATMENT ESTIMATES AND VARIANCES OF TREATMENT DIFFERENCES FOR VARIOUS CASES

Using the notation at the beginning of this section, we have

$$\hat{t}_i = Q_i/gEr, \tag{5}$$

$$\hat{t}_j^{(s)} = \frac{1}{Er}\left[Q_j^{(s)} + \frac{1}{c}\sum_{j=1}^{v-c} Q_j^{(s)} + \frac{1}{cg}\sum_{i=1}^{c} Q_i\right]. \tag{6}$$

Further note that, if $t'^{(s)}$ is the estimate of the jth regular treatment only using information from the sth experiment and if $t'_{i,s}$ has a similar meaning for the ith common treatment, then

$$\hat{t}_i = \frac{1}{g}\sum_{s=1}^{g} t'_{i,s} , \tag{7}$$

$$\hat{t}_j^{(s)} = t_j'^{(s)} + \frac{1}{c}\sum_{j=1}^{v-c} t_j'^{(s)} + \frac{1}{g}\sum_{i=1}^{c} \hat{t}_i , \tag{8}$$

or

$$\hat{t}_j^{(s)} = t_j'^{(s)} - \frac{1}{c}\sum_{i=1}^{c} t'_{i,s} + \frac{1}{g}\sum_{i=1}^{c} \hat{t}_i . \tag{8a}$$

Thus to find the best estimates, one need only estimate the treatments for each experiment seperately and use (7) and (8) or (8a) to obtain the final estimates. Formulae (8) and (8a) could be used with advantage when $(v - c) < c$ and $(v - c) > c$, respectively.

At this stage, it might be noted that the adjusted treatment sums of squares may also be obtained by the following alternative formula:

$$\text{Adjusted treatment S.S.} = \sum_{i=1}^{c} \hat{t}Q_i + \sum_{s=1}^{g}\sum_{j=1}^{v-c} \hat{t}_j'^{(s)}Q_j^{(s)}. \tag{9}$$

The variances of the various treatment differences are obtained as follows:

(1) For treatment estimates $\hat{t}_i$, $\hat{t}_{i'}$ belonging to the common group

$$\hat{V}_1 = \text{var}\,(\hat{t}_i - \hat{t}_{i'}) = 2s^2/gEr.$$

(2) For treatment estimates $\hat{t}_i$ and $\hat{t}_j^{(s)}$ where $\hat{t}_i$ is from common group and $\hat{t}_j$ is from regular group.

$$\hat{V}_2 = \text{var}\,(\hat{t}_i - \hat{t}_j^{(s)}) = \left\{\frac{1}{rE}\left[1 + \frac{1}{g} + \frac{1}{c} - \frac{1}{cg}\right]s^2\right\}.$$

(3) For treatment estimates $\hat{t}_j^{(s)}$ and $\hat{t}_{j'}^{(s)}$ where both of them belong to the regular group of the same experiment

$$\hat{V}_3 = {}_{\text{Var}}\,(\hat{t}_j^{(s)} - \hat{t}_{j'}^{(s)}) = 2s^2/rE.$$

(4) For treatment estimates $t_j^{(s)}$ and $t_{j'}^{(s')}$ where both of them belong to the regular group but are from different experiments

$$\hat{V}_4 = {}_{\text{Var}}\,(\hat{t}_j^{(s)} - \hat{t}_j^{(s')}) = \frac{2}{rE}\left(1 + \frac{1}{c}\right)s^2.$$

s^2 in all of the above formulae stands for the estimated error variance.

SPECIAL CASES

(a) There may be several BIB experiments with only one treatment common which might be a control. In this case $c = 1$.

Let t_1 denote the control. Then we have the following results.

I. *Adjusted treatment S. S.*

Sum of the adjusted treatment sums of squares from individual analyses

II. *Treatment estimates.*

$$\hat{t}_1 = \frac{1}{g} \sum_{s=1}^{g} t'_{1,s}\ ,$$

$$\hat{t}_j^{(s)} = t_j'^{(s)} - t'_{1,s} + \frac{1}{g}\,\hat{t}_1 \quad \text{based on (8a).}$$

III. *The variances of treatment differences.*

(1) This case does not arise

(2) $\hat{V}(\hat{t}_1 - \hat{t}_j^{(s)}) = 2s^2/rE.$

(3) $\hat{V}(\hat{t}_j^{(s)} - \hat{t}_{j'}^{(s)}) = 2s^2/rE.$

(4) $\hat{V}(\hat{t}_j^{(s)} - \hat{t}_j^{(s')}) = 4s^2/rE.$

The numbers here correspond to those in the preceding section.

(b) There may be several BIB experiments with the same treatments. In this case $c = v$. In other words, we want to combine several BIB experiments with identical parameter values and treatments, conducted at different places. In this case, we have results as follow:

I. *Adjusted treatments S.S.*

$$\frac{1}{rEg}\sum_{i=1}^{r} Q_i^2 .$$

II. *Treatment estimates.*

$$\hat{t}_i = \frac{1}{g}\sum_{s=1}^{g} t'_{i,s} \qquad i = 1, \cdots, c(=v).$$

III. *The variances of treatment differences.*

(1) $\text{var}\,(\hat{t}_i - \hat{t}_{i'}) = 2s^2/rEg, \qquad i \neq i', \qquad i, i' = 1 \cdots, v.$

(2), (3), (4) These cases do not arise.

(c) The most important special case is obtained when $b = r = \lambda$ and $k = v$. In this case BIB designs reduce to well known randomised complete block designs with c common treatments. The simplified analysis will remain the same with corresponding substitution.

(d) Cases (a) and (b) can be considered for (c).

Since (c) and (d) have been considered in detail by Gomes *et al* [1958], they will not be repeated here. However, we present the adjusted treatment sums of squares for the three situations since they are somewhat simpler than those given before.

Case (i), $c = 1$.

Adjusted treatment S. S. = Sum of treatment S. S. from the individual experiments.

Case (ii), $1 < c < v$.

Adjusted treatment S. S. = {Sum of treatment S. S. from the individual experiments} − {Interaction (common treatment × experiments) using treatment totals.}

Case (iii), $c = v$.

$$\text{Treatment S. S.} = \sum_{i=1}^{v} \frac{T_i^2}{rg} - \frac{G^2}{gvr}$$

where T_i is the total for the ith treatment over all experiments and G is the grand total for all experiments.

EXAMPLE

(Hypothetical problem purely to illustrate the procedure)

Two BIB experiments with the following parameter values are conducted and the combined analysis is required. The parameter values are

$$v = 7, \quad b = 7, \quad r = 4, \quad k = 4, \quad \lambda = 2, \quad g = 2 \text{ and } c = 3.$$

The following data are at our disposal.

TABLE 2

Analysis of Variance for the Individual Experiments

Source	Experiment 1		Experiment 2	
	d.f.	S.S.	d.f.	S.S.
Blocks (unadjusted)	6	342.71	6	63.21
Treatments (adjusted)	6	171.14	6	154.50
Residual	15	165.11	15	469.00

$$\begin{array}{llll} Q_{1,1} = 18.00, & Q_4^{(1)} = -1.00, & Q_{1,2} = 13.75, & Q_8^{(2)} = 4.50, \\ Q_{2,1} = 7.50, & Q_5^{(1)} = -9.75, & Q_{2,2} = -5.75, & Q_9^{(2)} = 0.50, \\ Q_{3,1} = -6.75, & Q_6^{(1)} = -8.75, & Q_{3,2} = -9.75, & Q_{10}^{(2)} = -8.25, \\ & Q_7^{(1)} = 0.75. & & Q_{11}^{(2)} = -5.00. \end{array}$$

Form, with $Q_{i,s}$ values, the following two-way table.

	t_1	t_2	t_3	Total
E_1	18.00	7.50	−6.75	18.75
E_2	13.75	−5.75	−9.75	−1.75
Total	31.75	1.75	−16.50	17.00

The (common treatment × experiments) interaction is found to be 31.2708.

Hence

$$\left.\begin{array}{l}\text{Treatment (adjusted) S. S.}\\ \text{for the combined analysis}\end{array}\right\}$$

$$= 171.14 + 154.50 - 4/2 \times 7\,(31.2708)$$
$$= 316.7055$$

The combined analysis is given below.

TABLE 3

COMBINED ANALYSIS OF VARIANCE

Source of variation	d.f.	S.S.	M.S.
Blocks (unadjusted)	13	425.375	32.7212
Treatments (adjusted)	10	316.708	31.6708
Residual	32	643.042	20.0951

With the use of the equations (7) and (8) we obtain the following treatment estimates for the combined analysis:

$$\begin{aligned} \hat{t}_1 &= 4.5357, & \hat{t}_2 &= 0.2500, & \hat{t}_3 &= -2.3571, \\ \hat{t}_4^{(1)} &= -1.2619, & \hat{t}_5^{(1)} &= -3.7619, & \hat{t}_6^{(1)} &= -3.4762, \\ \hat{t}_7^{(1)} &= -0.7619, & \hat{t}_8^{(2)} &= 2.2619, & \hat{t}_9^{(2)} &= 3.9762, \\ \hat{t}_{10}^{(2)} &= -1.3810, & \hat{t}_{11}^{(2)} &= -0.4524. \end{aligned}$$

Variances of treatment differences are

$$\begin{aligned} \hat{V}_1 &= 2s^2/Erg &&= 5.741, \\ \hat{V}_2 &= \frac{1}{Er}\left[1 + \frac{1}{g} + \frac{1}{c} - \frac{1}{cg}\right]s^2 &&= 9.569, \\ \hat{V}_3 &= 2s^2/Er &&= 22.966, \\ \hat{V}_4 &= \frac{2}{Er}\left(1 + \frac{1}{c}\right)s^2 &&= 15.310. \end{aligned}$$

ACKNOWLEDGMENTS

The author wishes to express his sincere thanks to G. S. Murty, Director, Tobacco Research, Central Tobacco Research Institute, Rajahmundry and V. N. Murty, Assistant Director, Central Statistical Organization, India for going through the manuscript and suggesting suitable modifications in the text.

The author also wishes to thank M. Subrahmanyam, Statistical Section, Central Tobacco Research Institute, Rajahmundry for indispensable help during the progress of this work.

REFERENCES

(1) Gomes, F. F. and Guimares, R. F. [1958]. Joint analysis of experiments in complete randomised blocks with some common treatments. *Biometrics 14*, 521–26.

(2) Rao, C. R. [1947]. General methods of analysis for incomplete block designs. *J. Amer. Stat. Assoc. 42*, 541–61.

(3) Roy, S. N. and Sar-han, A. E. *On inverting a class of patterned matrices.* Part I. Mimeographed notes. University of North Carolina, U.S.A.

GENERALIZED ASYMPTOTIC REGRESSION AND NON-LINEAR PATH ANALYSIS[1]

MALCOLM E. TURNER[2], ROBERT J. MONROE AND HENRY L. LUCAS, JR.

Department of Experimental Statistics,
North Carolina State College, Raleigh, North Carolina, U.S.A.

1. INTRODUCTION

It is a fundamental fact of statistical inference that the information contained in an analysis of experimental data is the sum of the a priori information built into the statistical model employed and the a posteriori information contained in the data itself. For this reason the biometrician must be concerned not only with the efficiency of his estimation procedures but also with the adequacy of his descriptive model. Much care must be taken to ensure that all relevant a priori information is utilized in the construction of the model. In some cases it is possible to derive rather sophisticated theoretical models on the basis of acquired knowledge and intelligent hypothesizing. These models are often conveniently found as solutions of differential equations. In other cases little more may be known than that the biological process in question is continuous. In this latter case one may resort to polynomial models where the degree of the polynomial is either found empirically or by prior consideration of the number of "bends" which one can reasonably assume to take place in the process being studied. In certain other cases it may be known that the process approaches some asymptotic value. This is especially true in those cases known as "growth" processes. The general ineptness of polynomial models for purposes of describing such asymptotic situations has been repeatedly pointed out, although polynomials in the reciprocals may sometimes be used conveniently.

Stevens [1951] and Pimentel-Gomes [1953], writing in this journal, have discussed inferential methods related to one form of transcendental asymptotic model, the so-called exponential model,

[1] Supported, in part, by the Institute of Statistics from the Revolving Research Fund initiated by a grant from the General Education Board for the purpose of stimulating basic methodological research.

[2] Present address: Department of Biophysics and Biometry, Medical College of Virginia, Richmond 19, Va., U. S. A.

$$y = \alpha + \beta e^{-x/\gamma} + \epsilon, \tag{1}$$

where x and y are observables; α, β and γ are unknown parameters to be estimated; and ϵ is a random error. The above mentioned writers give convenient computational procedures for finding maximum likelihood estimates when ϵ has a normal distribution with constant variance and the errors are uncorrelated one with another. Even when the errors are non-normal these same methods provide least square estimates of the parameters. The numerical devices used for finding the estimates are based on methods devised by Gauss and described in detail in the books by Whitaker and Robinson [1944] and Deming [1943]. The method of estimation, maximum likelihood, is due originally to Gauss and is discussed extensively by Fisher [1956] who is mainly responsible for our knowledge of the properties of the method. Both the numerical devices and the estimation method are reviewed in the recent work by Williams [1959].

Asymptotic models fall conveniently into either the class of transcendental models, including the exponential, or the class of rational models, including models such as the rectangular hyperbola and the inverse square law.

This paper considers the general class of rational models and a special sub-class of transcendental models that may be regarded as a generalization of the exponential model. Appropriate techniques for estimation and for determination of asymptotic confidence limits, similar to those given by Stevens for the exponential, are provided. An alternative "parabolic method" of constructing confidence limits, valid for small samples, is also given.

In many practical biological situations a number of variables are related one with another in complex causal networks. Wright [1918, 1921] devised methods, which he termed path analysis, for analysizing such networks when the relationships between the variables are linear. Turner and Stevens [1959] have reviewed these methods for the important case when some of the primary variables may be "fixed" or without probability distributions. The present paper considers some aspects of such situations when the relationships between variables are of the transcendental variety rather than linear.

All of the models considered in this paper are non-linear in some of the unknown parameters. This implies that sufficient estimates and hence fiducial limits are not available.

In addition the maximum likelihood estimates, though consistent and efficient, will be generally biassed, will not have normal distributions for small samples and will generally not have explicit forms, thus requiring iterative solution methods. This is the price one has to pay for more adequate symbolic description of the process being studied. The gain in meaningfullness is generally well worth even this heavy toll.

2. RATIONAL REGRESSION

Let us denote a particular p-th degree polynomial by $P_p(x)$ and a particular q-th degree polynomial by $Q_q(x)$ and then let us consider the regression model arising from the rational function,

$$y = [P_p(x)/Q_q(x)] + \epsilon, \tag{2}$$

where again x and y are observables and ϵ is a random error assumed to have a normal distribution with zero expectation and constant variance σ^2. The errors are also assumed to be uncorrelated one with another and with the values of x. At least some of the coefficients in the polynomials are unknown parameters to be estimated.

The numerical device for estimation due to Gauss, which was alluded to in the previous section, involves expansion of $y - \epsilon$ about "trial values" of the unknown parameters in a Taylor's series. The method is sketched in Section 4 and is there applied to the models discussed in Section 3. However, a simpler approach is available for the rational model (2).

Suppose that $Q_q(x) = 1 + \beta_1 x + \cdots + \beta_q x^q$, so that (2) may be written: $y = [P_p(x)/(1 + \beta_1 x + \cdots + \beta_q x^q)] + \epsilon$. If we multiply through by $Q_q(x)$ and rearrange we get

$$y = P_p(x) - \beta_1 xy - \cdots - \beta_q x^q y + Q\epsilon. \tag{3}$$

Now (3) is an ordinary multiple linear regression model except for the factor Q in the error term. Thus weighted least squares could be used for estimating the unknown parameters if the weights Q^{-2} were known. A simple plan is to take trial values for the parameters $\beta_1 , \cdots , \beta_q$ and compute provisional weights. Then (3) is fitted by least squares giving better estimates of the β's as well as provisional estimates for $\alpha_0 , \alpha_1 , \cdots , \alpha_p$ occurring in $P_p(x)$. Using the improved estimates for the β's, better weights are determined and the fitting is repeated. The process is thus iterated through several cycles until stable values of the estimates are obtained. In many cases it is sufficient to use constant weights ($Q^{-2} = 1$) in the first cycle of iteration, thus not requiring trial values.

Asymptotic variances and confidence limits are found in the usual way from the final least square fit. Thus if for n observations we have

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p & x_1y_1 & x_1^2y_1 & \cdots & x_1^qy_1 \\ 1 & x_2 & x_2^2 & \cdots & x_2^p & x_2y_2 & x_2^2y_2 & \cdots & x_2^qy_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p & x_ny_n & x_n^2y_n & \cdots & x_n^qy_1 \end{bmatrix},$$

$$\mathbf{B}' = (\alpha_0 \ \alpha_1 \ \cdots \ \alpha_p \ \beta_1 \ \cdots \ \beta_q),$$

$$\mathbf{Y}' = (y_1 \ y_2 \ \cdots \ y_n),$$

$$\mathbf{W} = \begin{bmatrix} Q_1^{-2} & 0 & \cdots & 0 \\ 0 & Q_2^{-2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & Q_n^{-2} \end{bmatrix},$$

then the final estimates after iterating are

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \tag{4}$$

and the asymptotic variances and covariances are given by

$$\text{Cov}\,(\hat{\mathbf{B}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}s^2, \tag{5}$$

where s^2, the estimate of σ^2, is provided by

$$s^2 = (\mathbf{Y}'\mathbf{W}\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{W}\mathbf{Y})/(n - p - q - 1). \tag{6}$$

If the error in (2) depends upon x in a known way, only a slight modification of the procedure is required. Suppose that $\epsilon_x = f(x) \cdot \epsilon$. Then the error term in (3) will be $fQ\epsilon$ and the weight matrix **W** will be diagonal with elements $(f_iQ_i)^{-2}$ instead of $(Q_i)^{-2}$. Formulas (4), (5), and (6) will remain the same.

Computational details for those not familiar with linear regression methods are given by Williams [1959]. Let the elements of $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ be denoted by c_{ij}; then the individual asymptotic confidence limits for each coefficient are

$$\begin{aligned} \hat{\alpha}_i - t_\pi\sqrt{c_{ii}s^2} \le \alpha_i \le \hat{\alpha}_i + t_\pi\sqrt{c_{ii}s^2}; \qquad & i = 0, 1, \cdots, p, \\ \hat{\beta}_i - t_\pi\sqrt{c_{ii}s^2} \le \beta_i \le \hat{\beta}_i + t_\pi\sqrt{c_{ii}s^2}; \qquad & i = 1, 2, \cdots, q, \end{aligned} \tag{7}$$

where t_π is a Student's t at an assigned confidence level $(1 - \pi)$ with $n - p - q - 1$ degrees of freedom.

As an example consider the case of $p = 0$ and $q = 1$, the case of the rectangular hyperbola. We have

$$y = [\alpha_0/(1 + \beta_1 x)] + \epsilon \tag{8}$$

with asymptotes $x = -1/\beta_1$ and $y = 0$. Corresponding to (3) we have the regression equation $y = \alpha_0 - \beta_1 xy + Q\epsilon$, where $Q = 1 + \beta_1 x$.

A graphical estimate of the slope $-\beta_1$ can be obtained by plotting y against xy and provisional weights computed from $w = (1 + \beta_1 x)^{-2}$. Then using these provisional weights we compute

$$\hat{\beta}_1 = -\frac{\sum w \sum wxy^2 - \sum wxy \sum w}{\sum w \sum wx^2y^2 - (\sum wxy)^2} = -\frac{\sum w\, S_{xyy}}{\sum w\, S_{xxyy}} \tag{9}$$

and obtain better estimates of the weights. After stability is reached we find

$$\hat{\alpha}_0 = (\sum wy + \hat{\beta}_1 \sum wxy)/\sum w \tag{10}$$

and the estimate of the variance σ^2,

$$s^2 = \frac{1}{n-2}\left\{\frac{\sum w \sum wy^2 - (\sum wy)^2}{\sum w} - \frac{S_{xyy}^2}{S_{xxyy}}\right\}. \tag{11}$$

We also obtain estimates of the asymptotic variances for a_0 and b_1 :

$$\begin{aligned} s_{a_0}^2 &= s^2\left\{\frac{1}{\sum w} + \frac{(\sum wxy)^2}{(\sum w)^2\, S_{xxyy}}\right\} \\ s_{b_1}^2 &= s^2/S_{xxyy} \end{aligned} \tag{12}$$

from which asymptotic confidence limits may be calculated.

3. TRANSCENDENTAL REGRESSION: THE SINGLE PROCESS LAW

The rational models discussed in the previous section may often be used to conveniently describe an asymptotic process. However, when more than just a few terms in the polynomials are required for adequate description of the process, the rational models may suffer from three deficiencies: (1) physical or biological meaning of the coefficients may be lost, (2) annoying "bumps" tend to appear in the fitted curve which have no reality in the actual process, and (3) excessive numerical computation may be required due to the number of parameters which need to be estimated.

Certain transcendental models may provide a more parsimonious description of an asymptotic process and yet preserve considerable flexibility of form. By "parsimony" is meant the ability to describe adequately with a small number of parameters. This may be accomplished by supplying the a priori information that the curve is regular in some way or another. For example, we may specify that the curve is monotonically increasing or decreasing throughout its course or that certain of the derivatives have this monotonicity property. We may then avoid the undesired "bumps" and make possible simpler computations by using this information. In addition, since solutions of most differential equations are in terms of transcendental functions, we may wish to seek a model which is related to a class of differential equations, thus making interpretation of the parameters more meaningful.

One simple class of transcendental regression models has been called by Turner [1959] the *single process law.* Consider the differential equation

$$d\eta/d\xi = \delta(\eta - \alpha)/(\xi - \gamma\ \delta). \tag{13}$$

Integrating we get

$$\begin{aligned} &\eta = \alpha + \beta(\xi - \gamma\ \delta)^{\delta}, \\ \text{or} \qquad &\eta = \alpha + \beta(\xi - \mu)^{\delta}, \qquad \mu = \gamma\ \delta. \end{aligned} \tag{14}$$

Now let us suppose that the observables x and y are defined by the error equations $x = \xi$, and $y = \eta + \epsilon$, where ϵ is a random error defined as before. Thus the model corresponding to the single process law (14) is

$$y = \alpha + \beta(x - \gamma\ \delta)^{\delta} + \epsilon. \tag{15}$$

For convenience we will always take $x - \gamma\delta > 0$.

The proportionality constant δ in (13) determines the nature of the process. When δ is a positive integer, we have a polynomial process; however, the model (15) is non-linear when this integer is greater than two. When δ is a negative integer, then we have a rational process, a special case of the processes studied in the previous section. When δ is non-integral, the process is transcendental. In particular, when δ approaches infinity, either positively or negatively, the model (15) approaches the exponential model (1), $y = \alpha + \beta e^{-x/\gamma} + \epsilon$.

An interesting series of models, most of which are commonly used in curve fitting, is obtained when $2/\delta = -4, -3, -2, -1, 0, +1, +2$. These cases correspond to the *inverse square root law*, the *inverse two-thirds law*, the *rectangular hyperbola*, the *inverse square law*, the *exponential*, the *parabola*, and the *straight line*, respectively. Thus all of these are special cases of a continuously varying family of curves (See Figure 1). We observe that when $2/\delta$ is negative the curves have two asymptotes, that when $2/\delta = 0$ there is but one asymptote, and that when $2/\delta$ is positive there is no asymptote. Thus the exponential may be thought of as a "transition" curve between those curves possessing two asymptotes and those possessing none. Without loss of generality we may consider only the case of decreasing curves, commonly referred to as "extinction curves". Several useful constants for the

TABLE 1
CONSTANTS FOR SPECIAL EXTINCTION CURVES[a]

δ^{-1}	Name of Curve	δ	β	γ	Asymptote	Initial level	ξ^{b}_{25}	ξ_{50} "half-life"	ξ_{75}	ξ^{c}_{100} "point of extinction"	$\lambda^{d} = \xi_{75}/\xi_{50}$	Initial Velocity
1.0	Straight Line	1	<0	>0	—	$-\beta\gamma$	$\gamma/4$	$\gamma/2$	$3\gamma/4$	γ	1.5	β
.5	Parabola	2	>0	>0	—	$4\beta\gamma^2$	$(2 - \sqrt{3})\gamma$	$(2 - \sqrt{2})\gamma$	γ	2γ	1.7071	$-4\beta\gamma$
0	Exponential	∞	>0	>0	—	β	$\left(\log \frac{4}{3}\right)\gamma$	$(\log 2)\gamma$	$(\log 4)\gamma$	—	2.0	$-\frac{\beta}{\gamma}$
$-$ .5	Inverse Square Law	-2	>0	>0	-2γ	$\frac{\beta}{4\gamma^2}$	$\left(\frac{4}{3}\sqrt{3} - 2\right)\gamma$	$(2\sqrt{2} - 2)\gamma$	2γ	—	2.4142	$-\frac{\beta}{4\gamma^3}$
-1.0	Rectangular Hyperbola	-1	>0	>0	$-\gamma$	$\frac{\beta}{\gamma}$	$\gamma/3$	γ	3γ	—	3.0	$-\frac{\beta}{\gamma^2}$
-1.5		$-2/3$	>0	>0	$-\frac{2\gamma}{3}$	$\frac{\beta}{\sqrt[3]{4\gamma^2/9}}$	$\frac{2}{3}\left(\frac{8\sqrt{3}}{3} - 3\right)\gamma$	$\frac{2}{3}(2\sqrt{2} - 1)\gamma$	$\frac{14}{3}\gamma$	—	3.8284	$-\frac{\beta}{\gamma\sqrt[3]{4\gamma^2/9}}$
-2.0	Inverse Square Root Law	$-1/2$	>0	>0	$-\frac{\gamma}{2}$	$\frac{\beta}{\sqrt{\gamma/2}}$	$7\gamma/18$	$3\gamma/2$	$15\gamma/2$	—	5.0	$-\frac{\beta}{\gamma\sqrt{\gamma/2}}$

[a] The curves all have the following characteristics in common:
(1) Finite, positive initial value of $\eta(\xi = 0)$
(2) $d\eta/d\xi < 0$ for all values of ξ
(3) $d^2\eta/d\xi^2 \geq 0$ for all values of ξ
(4) min $\eta = \alpha$ except for the straight line values for which we will ignore beyond the point $(\xi = \gamma, \eta = \alpha)$.

[b] $\xi_p = \gamma\,\delta(1 - q^{1/\delta}), \quad p + q = 1.$

[c] $\xi_{100} = \delta\gamma, \quad \delta > 0.$

[d] $\lambda = \frac{(2^{1/\delta})^2 - 1}{(2^{1/\delta})^2 - 2^{1/\delta}}, \quad \delta = -\frac{\log 2}{\log(\lambda - 1)}; \text{ for } -\infty < \frac{1}{\delta} < +\infty, \quad \lambda > 1.$

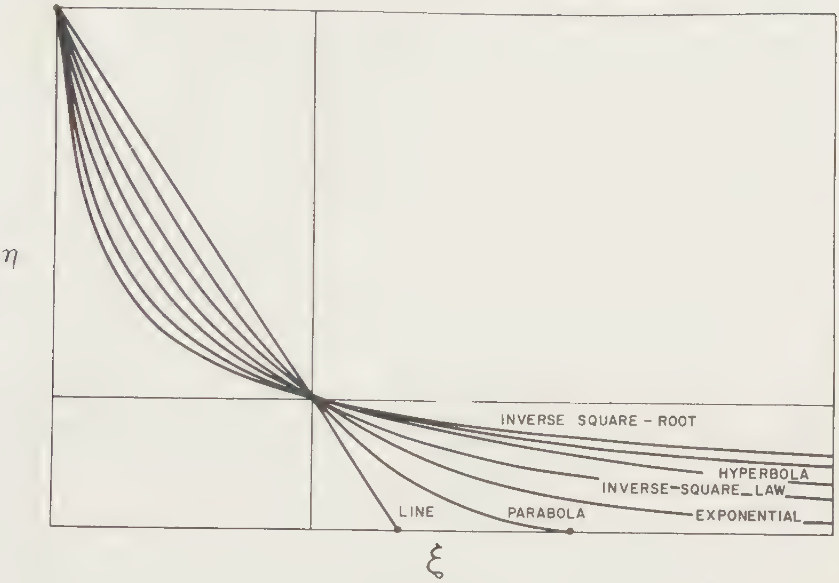


FIGURE 1
SPECIAL EXTINCTION CURVES

special cases enumerated above are summarized in Table 1. In this table the minimum level is taken in all cases to be α and the initial level is measured from this minimum value.

Table 1 includes the *three quarters life* and the *half-life.* The ratio of the two provides a convenient means of estimating δ graphically. From equation (14) we determine the initial value of η to be $\eta_0 = \alpha + \beta(-\gamma\delta)^{\delta}$ and the minimum value of η to be $\eta_{\min} = \alpha$. The value of ξ corresponding to $(\eta_0 - \alpha)/2$ is the *half-life* and is given by $\xi_{50} = \gamma\delta(1 - 2^{-1/\delta})$. Similarly we obtain the *three-quarters life* $\xi_{75} = \gamma\delta(1 - 4^{-1/\delta})$. The ratio is then $\lambda = \xi_{75}/\xi_{50} = (1 - 4^{-1/\delta})/(1 - 2^{-1/\delta})$, or, inverting, we have

$$\delta^{-1} = -\log_2 (\lambda - 1). \tag{16}$$

The parameter λ is given for each of the special cases in Table 1. and the relationship (16) is charted for all cases in Figure 2.

A more general criterion based upon any two percentile-lives is readily found to be

$$\lambda(p_1 , p_2) = \xi_{p_1}/\xi_{p_2} = (1 - q_1^{1/\delta})/(1 - q_2^{1/\delta}), \tag{17}$$

where $p_1 + q_1 = p_2 + q_2 = 1$.

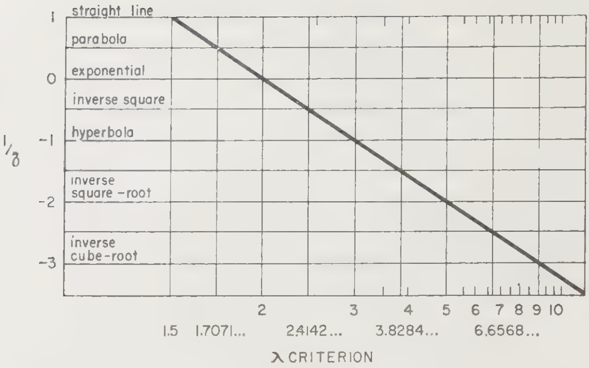


FIGURE 2

Relationship Between the λ Criteron and the Exponential Parameter δ

4. THE GAUSSIAN ITERANT

Gauss appears to have been the first to use the Newton-Raphson technique for purposes of finding least squares estimates of non-linear parameters. See Whitaker and Robinson [1944]. In the case of regression with normally distributed errors, Gauss' technique is equivalent to the important *method of scoring* due to Fisher. See Fisher's recent discussion [1956] of the subject.

The technique may be conveniently described in terms of the total derivative. Let us suppose that we have the general regression model

$$y = f(\boldsymbol{\beta}, x) + \epsilon$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters. The fitted "prediction" equation is, in terms of estimates $\mathbf{b}$,

$$\hat{y} = f(\mathbf{b}, x). \tag{18}$$

By the rule for the total derivative we get $d\hat{y} = \hat{y}_1\, db_1 + \hat{y}_2\, db_2 + \cdots$, where $\hat{y}_i = \partial\hat{y}/\partial b_i$. Then to a first approximation

$$\Delta\hat{y} \doteq \hat{y}_{10}\, \Delta b_1 + \hat{y}_{20}\, \Delta b_2 + \cdots,$$

where $\Delta\hat{y} = y - y_0$, $\Delta b_1 = b_1 - b_{10}$, $\Delta b_2 = b_2 - b_{20}$, $\cdots$. The zero subscript indicates that trial values for the undetermined estimates have been inserted. Thus, we get $\hat{y} \doteq \hat{y}_0 + \Delta b_1 \hat{y}_{10} + \Delta b_2 \hat{y}_{20} + \cdots$, and finally

$$y \doteq \hat{y}_0 + \Delta b_1 \hat{y}_{10} + \Delta b_2 \hat{y}_{20} + \cdots + e. \tag{19}$$

The symbol e represents the estimate of the error ϵ. Now (19) is an ordinary multiple linear regression equation and estimates may be found as before. From the linear regression estimates may be found improved values for the original non-linear regression estimates. Thus, the process is iterated finding successive improvements. It may be seen from (19) that trial values for estimates which are linear in (18) are not needed to prime the iterative process. This point was noted by Stevens [1951] and will be illustrated in the next section. Refer to Deming [1943] for a fuller discussion of this technique of estimation.

5. MAXIMUM LIKELIHOOD ESTIMATION AND ASYMPTOTIC CONFIDENCE LIMITS FOR THE SINGLE PROCESS LAW

Applying the method of the previous section to the single process law we obtain the linearized regression equation:

$$\begin{aligned} y \doteq a + b(x - m_0)^{d_0} - b\, d_0(x - m_0)^{d_0 - 1}(m - m_0) \\ + b(x - m_0)^{d_0} \ln (x - m_0)(d - d_0) + e, \end{aligned} \tag{20}$$

where a_0, b_0, m_0, and d_0 are trial values and $m = cd$. Note that (20) is linear in a and b and the corrections $(m - m_0)$ and $(d - d_0)$. Thus provisional estimates of a and b and improved estimates of m and d can be found by fitting (20) by least squares. Iteration provides the maximum likelihood estimates. We notice that trial values of the linear coefficients a and b have been eliminated and are thus not required.

For convenience we write

$$\begin{aligned} B_0 &= a, & Z_0 &= 1, \\ B_1 &= b, & Z_1 &= (x - m_0)^{d_0}, \\ B_2 &= -b\, d_0(m - m_0), & Z_2 &= (x - m_0)^{d_0 - 1}, \\ B_3 &= b(d - d_0), & Z_3 &= (x - m_0)^{d_0} \ln (x - m_0). \end{aligned} \tag{21}$$

Then (20) may be written

$$y \doteq B_0 + B_1 Z_1 + B_2 Z_2 + B_3 Z_3 + e.$$

We have then to find estimates of B_0, B_1, B_2, and B_3. These are found by multiple regression methods as follows. Let

$$\mathbf{Z} = \begin{bmatrix} 1 & Z_{11} & Z_{12} & Z_{13} \\ 1 & Z_{21} & Z_{22} & Z_{23} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & Z_{n1} & Z_{n2} & Z_{n3} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \\ B_3 \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}.$$

The normal equations are $\mathbf{Z'ZB} = \mathbf{Z'Y}$ and, if $\mathbf{Z'Z}$ is non-singular, the vector of improved estimates is given by

$$\hat{\mathbf{B}} = (\mathbf{Z'Z})^{-1}\mathbf{Z'Y}. \tag{22}$$

Whenever the error depends upon x in a known way, (22) must be replaced by $\hat{\mathbf{B}} = (\mathbf{Z'WZ})^{-1}\mathbf{Z'WY}$ where $\mathbf{W}$ is a diagonal matrix of weights as in Section 2. From the results of (22) we obtain improved values of m and d by using the relations (21). Thus

$$\begin{aligned} m &= m_0 - \hat{B}_2/d_0\hat{B}_1 , \\ d &= d_0 + \hat{B}_3/\hat{B}_1 . \end{aligned} \tag{23}$$

These improved values replace m_0 and d_0 in (21) and the solution (22) is found again, giving still better values. The process is iterated until constant values of m and d are found to the desired degree of precision. After stable values are obtained we have

$$a = B_0 , \qquad b = B_1 , \qquad c = m/d, \tag{24}$$

and the estimated curve is given by

$$\hat{y} = a + b(x - m)^d. \tag{25}$$

If $\mu = \gamma\delta$ is known a priori, then the third row and column of $\mathbf{Z'Z}$, the third element of $\mathbf{B}$ and the third element of $\mathbf{Z'Y}$ are deleted. Otherwise, the iterative process is the same as before. Application of this case to a biomathematical model of running records has been made by Turner [1959]. Similarly, when δ is known a priori, the fourth row and column of $\mathbf{Z'Z}$ and the third elements of $\mathbf{B}$ and $\mathbf{Z'Y}$ are deleted. If, as is often the case in extinction curves, the minimum value α is known a priori then y is replaced by $y' = y - \alpha$ and the first row and column of $\mathbf{Z'Z}$ and the first elements of $\mathbf{B}$ and $\mathbf{Z'Y}$ are deleted. Thus, we see that one estimation procedure suffices for the general problem and several special cases as well. However, in the case that δ is a known integer, we may prefer to use the methods of Section 2. If δ is known to be infinite, the methods of Stevens and Pimentel-Gomes referred to earlier should be used.

By examination of the definition of the Z's, equations (21), we see that $\mathbf{Z'Z}$ is singular if $\delta = 0, 1,$ or $\pm\infty$. Therefore, the method of this section may be expected to break down in the neighborhood of these points. Trial and error minimization of the residual sum of squares, $S = \sum [y - a - b(x - m)^d]^2$, may then be resorted to, perhaps utilizing some interpolation method such as that proposed by Will [1936]. Values of m and d may be selected and S minimized by ordinary straight line regression techniques. Thus for the selected values m_1 and d_1 we may regress y on $X = (x - m_1)^{d_1}$, obtaining the minimum S for these values: $S_{11} = \sum (y - \bar{y})^2 - \{[\sum X(y - \bar{y})]^2/\sum (X - \bar{X})^2\}$. Similarly, values S_{12} for m_1 and d_2, S_{22} for m_2 and d_2, et cetera, may be found and the minimum located.

Alternatively, we may choose selected values of δ only (or of μ only) and use the deleted normal equations to minimize with respect to α, β, and μ (or α, β, and δ). Then simple one-dimensional interpolation will locate the minimum value of S.

When $\mathbf{Z'Z}$ is only moderately ill-conditioned, we may improve convergence by damping oscillations. Wegstein [1958] recommends using in the $i + 1$-th iteration (in place of m_i and d_i) the modified values

$$\begin{aligned} m_i^* &= um_{i-1} + vm_i\ , \\ d_i^* &= u\,d_{i-1} + v\,d_i\ . \end{aligned} \tag{26}$$

The following data taken from a sampling experiment [Turner, 1959] will illustrate the iterative process.

x	y
1	162.47
2	108.10
3	81.12
4	66.11
5	55.06
6	49.37
7	39.72
8	33.81
9	30.90
10	30.73

Trial values $m_0 = -1.00$ and $d_0 = -1.00$ were tried and the results below were obtained for successive iterations.

Iteration	m	d
0	−1.00	−1.00
1	− .23	− .56
2	− .38	− .69
3	− .43	− .67
4	− .44	− .67

Asymptotic confidence limits are easily found. After the final iteration the asymptotically unbiassed estimate of σ^2 is given by

$$s^2 = (\sum y^2 - B_0 \sum y - B_1 \sum Z_1 y - B_2 \sum Z_2 y - B_3 \sum Z_3 y)/(n - 4). \qquad (27)$$

Let us denote by c_{ij} the element in the ith row and jth column of $(\mathbf{Z}'\mathbf{Z})^{-1}$. It has been well known since Gauss that the variance σ_f^2 of any function of the B's (f say) is asymptotically given by

$$\sigma_f^2 = \sigma^2 \sum_{i=0}^{3} \sum_{j=0}^{3} c_{ij} f_i f_j ,$$

where $f_i = \partial f / \partial B_i$ and $f_j = \partial f / \partial B_j$. We may use as the estimated variance of the function the quantity

$$s_f^2 = s^2 \sum_{i=0}^{3} \sum_{j=0}^{3} c_{ij} f_i f_j . \qquad (28)$$

Then from (23) and (24) using (28) we find the estimated asymptotic variances for a, b, m, d, and c to be

$$\begin{aligned} s_a^2 &= c_{00} s^2, \qquad s_m^2 = c_{22} s^2 / b^2 d^2, \\ s_b^2 &= c_{11} s^2, \qquad s_d^2 = c_{33} s^2 / b^2, \\ s_c^2 &= [(c_{22}/b^2 d^2 m^2) - (2c_{23}/b^2 d^3) + (m^2 c_{33}/b^2 d^4)] s^2. \end{aligned} \qquad (29)$$

Finally, for y, using (25) and (28), we find the asymptotic variance

$$s_{\hat{y}}^2 = s^2 \sum_{i=0}^{3} \sum_{j=0}^{3} c_{ij} Z_i Z_j . \qquad (30)$$

From (29) and (30) we can construct asymptotic confidence limits. We get

$$\begin{aligned} a - t_\pi s_a \le \alpha &\le a + t_\pi s_a , \\ b - t_\pi s_b \le \beta &\le b + t_\pi s_b , \\ m - t_\pi s_m \le \mu &\le m + t_\pi s_m , \\ d - t_\pi s_d \le \delta &\le d + t_\pi s_d , \\ c - t_\pi s_c \le \gamma &\le c + t_\pi s_c , \\ \hat{y} - t_\pi s_{\hat{y}} \le E(y) &\le \hat{y} + t_\pi s_{\hat{y}} , \end{aligned} \qquad (31)$$

with $(1 - \pi)$ confidence each where t_π is Student's t for the π level of significance with $n - 4$ degrees of freedom. The last confidence statement provides confidence limits for the true value of $\eta = E(y)$

for a given x. See Deming [1943] for detailed discussion of these methods for any model. It must be emphasized that (31) provides individual confidence statements at the assigned $(1 - \pi)$ level of confidence. Of course, the joint confidence level is much less than $(1 - \pi)$.

These confidence statements provide a ready guide to the tenability of various hypotheses about the unknown parameters in the single process law. If it is specifically hypothesized on theoretical grounds that some particular parameter takes a given value (e.g., $H_0 : \delta = \delta_0$), then we may test the concordance of the data with this hypothesis by seeing whether or not the hypothesized value lies within the confidence interval. If the hypothesized value is in fact tenable, we may wish to refit the regression equation after making the proper deletions for a known a priori parameter. On the other hand, if there is no theoretical reason for making such a hypothesis, it does not seem wise to recompute the curve for some convenient value of the parameter lying within the confidence interval. Such practice will lead to a false sense of confidence in that narrower confidence limits will result from fixing the parameter.

6. PARABOLIC CONFIDENCE LIMITS FOR SMALL SAMPLES

The asymptotic confidence limits described in the foregoing section may be expected to give adequate results for large samples or even for quite modest samples if the error rate is small. For example, in a recent sampling investigation [Turner, 1959], it was found that, with a sample size of ten equally spaced points covering the major portion of the curve and with a coefficient of variation of two percent at the center of the curve, the sampling variation of d, the estimate of δ, was quite accurately described by the asymptotic distribution. However, when the error rate is much larger, it may be supposed that the asymptotic distribution will not suffice for construction of tests or confidence limits.

We will now describe a method of constructing an exact test and corresponding confidence limits based upon a class of "parabolic" alternative hypotheses.

We consider the null hypothesis

$$H_0 : \mu - \mu_0 = \delta - \delta_0 = 0. \tag{32}$$

Thus, when the null hypothesis H_0 is true, we have the model

$$y_i = \alpha + \beta(x_i - \mu_0)^{\delta_0} + \epsilon_i \ , \tag{33}$$

where $i = 1, 2, \cdots, n$ for sample size n.

Let $Z_{i0} = (x_i - \mu_0)^{\delta_0}$; then the model (33) for H_0 true is written

$$y_i = \alpha + \beta Z_{i0} + \epsilon_i ,$$

the model for a straight line.

Now, if H_0 is not true, we could write

$$y_i = f(Z_{i0}) + \epsilon_i . \tag{34}$$

Expanding (34) in a Maclaurin series and preserving terms through the quadratic, we get

$$y_i \doteqdot B_0 + B_1 Z_{i0} + B_2 Z_{i0}^2 + \epsilon_i .$$

If we suppose that coefficients of terms beyond the quadratic are negligible, then the null hypothesis,

$$H_0' : B_2 = 0,$$

is equivalent to (32). An exact test of H_0' is well known and is easily performed.

Let

$$\begin{aligned}
S_{11} &= \sum Z_{i0}^2 - [(\sum Z_{i0})^2/n], \\
S_{12} &= \sum Z_{i0}^3 - [(\sum Z_{i0})(\sum Z_{i0}^2)/n], \\
S_{1y} &= \sum Z_{i0} y_i - [(\sum Z_{i0})(\sum y_i)/n], \\
S_{22} &= \sum Z_{i0}^4 - [(\sum Z_{i0}^2)^2/n], \\
S_{2y} &= \sum Z_{i0}^2 y_i - [(\sum Z_{i0}^2)(\sum y_i)/n], \\
S_{yy} &= \sum y_i^2 - [(\sum y_i)^2/n],
\end{aligned} \tag{35}$$

where all summations are over $i = 1, 2, \cdots, n$. Then we have estimates

$$\begin{aligned}
\hat{B}_1 &= (S_{22}S_{1y} - S_{12}S_{2y})/(S_{11}S_{22} - S_{12}^2), \\
\hat{B}_2 &= (S_{11}S_{2y} - S_{12}S_{1y})/(S_{11}S_{22} - S_{12}^2).
\end{aligned} \tag{36}$$

Then the residual sum of squares about the fitted quadratic curve is given by

$$SSE = S_{yy} - \hat{B}_1 S_{1y} - \hat{B}_2 S_{2y} . \tag{37}$$

Similarly, the residual sum of squares about the straight line is given by

$$SSE' = S_{yy} - (S_{1y}^2/S_{11}). \tag{38}$$

Then an exact F-test for $H_0' : B_2 = 0$, which is approximately equivalent to $H_0 : \mu - \mu_0 = \delta - \delta_0 = 0$, is performed by calculating

$$F = (n - 3)(SSE' - SSE)/SSE. \tag{39}$$

Then if $F \geq F_\pi$, the tabular value of F with 1 and $n - 3$ degrees of freedom, the null hypothesis is rejected at the π level of significance.

A joint confidence region, with $(1 - \pi)$ confidence, for μ and δ (and hence γ and δ) may be found by determining those values of μ_0 and δ_0 which ensure that $F \leq F_\pi$. This may be done most conveniently, perhaps, by calculating F for various selected values of μ_0 and δ_0 and then interpolating to find the confidence contours. A numerical example when μ is known a priori has been given by Turner [1959].

In some cases it may be necessary to retain higher-power terms in the Maclaurin expansion. Since then a wider class of alternative hypotheses would be admitted, the power of the test against any specific hypothesis would in general be lower.

When observations are replicated for various values of x, the numerator of (39) may be tested against the *within replicate* sum of squares to give a more powerful test. If the *within replicate* sum of squares is denoted SSE^*, we have the test

$$F = n(r - 1)(SSE' - SSE)/SSE^*, \tag{40}$$

with 1 and $n(r - 1)$ degrees of freedom, where n is the number of points on the curve as before and r is the number of replicates. A confidence region may be found corresponding to this test just as before.

A test similar in concept to the parabolic test has been proposed by Williams [1959] and is discussed in connection with the single process law by Turner [1959]. Williams' test is not likely to be as powerful as the parabolic test whenever there is more than a single non-linear parameter, as in the present case. If either μ or δ is known a priori then Williams' test should be considered as an alternative to the parabolic test.

Hotelling [1939] has proposed a method of constructing exact tests which could be applied to the construction of a test of the hypothesis $\beta = 0$ in the single process law. However β is the constant of integration and is generally of little interest since the character of the process is not determined by this constant.

7. NON-LINEAR PATH REGRESSION AND THE SINGLE PROCESS LAW

Let us suppose that we have p "secondary" variables which are causally related to one another and to a set of q "primary" or "controllable" variables. We imagine that the q primary variables (denoted $\xi_a, \xi_b, \cdots, \xi_q$) have values which can be chosen at will by the experimenter. In particular, it is possible that they are randomly chosen but this is generally not so in laboratory experimentation. The p

secondary variables (denoted $\eta_1, \eta_2, \cdots, \eta_p$) are then supposed to be completely determined. The rates of change of one variable with respect to another are determined by a matrix of partial derivatives

$$\mathbf{H}' = \left[\begin{array}{cccc} \eta_{1a} & \eta_{2a} & \cdots & \eta_{pa} \\ \eta_{1b} & \eta_{2b} & \cdots & \eta_{pb} \\ \cdots & \cdots & \cdots & \cdots \\ \eta_{1q} & \eta_{2q} & \cdots & \eta_{pq} \\ \hline \eta_{11} & \eta_{21} & \cdots & \eta_{p1} \\ \eta_{12} & \eta_{22} & \cdots & \eta_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ \eta_{1p} & \eta_{2p} & \cdots & \eta_{pp} \end{array}\right] = \begin{bmatrix} \mathbf{H}'_\xi \\ \mathbf{H}'_\eta \end{bmatrix}. \tag{41}$$

When $\eta_{ij} = \alpha_{ij}$, a constant, then the relations between all of the variables are linear. A recent review of linear path regression analysis is that of Turner and Stevens [1959]. See also Turner [1959]. The method of path regression analysis, originally described and named by Wright [1921], has been applied only in the case of linear relations, with one exception [Tukey, 1954]. The method, however, only attains its full power when the relations between variables are described by non-linear equations. This is partly what Tukey meant when in his critique [Tukey, 1954] he said that classical path coefficients are "very good, but they don't go far enough". For example, the important field of chemical reaction kinetics requires a path analysis wherein the relations between variables are governed by the *law of mass action.*

In many cases we may regard η_{ij} to be a function of η_i and ξ_j (or η_j) only. In particular, we may consider that the relations between variables follow the single process law.

Let a matrix of rate coefficients be

$$\mathbf{\Delta}' = \left[\begin{array}{cccc} \delta_{1a} & \delta_{2a} & \cdots & \delta_{pa} \\ \delta_{1b} & \delta_{2b} & \cdots & \delta_{pb} \\ \cdots & \cdots & \cdots & \cdots \\ \delta_{1q} & \delta_{2q} & \cdots & \delta_{pq} \\ \hline 1 & \delta_{21} & \cdots & \delta_{p1} \\ \delta_{12} & 1 & \cdots & \delta_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ \delta_{1p} & \delta_{2p} & \cdots & 1 \end{array}\right] = \begin{bmatrix} \mathbf{\Delta}'_\xi \\ \mathbf{\Delta}'_\eta \end{bmatrix}$$

and diagonal matrices be

$$\mathbf{D}_\xi = \begin{bmatrix} \xi_a & 0 & \cdots & 0 \\ 0 & \xi_b & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \xi_q \end{bmatrix}, \qquad \mathbf{D}_\mu = \begin{bmatrix} \mu_a & 0 & \cdots & 0 \\ 0 & \mu_b & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mu_q \end{bmatrix},$$

$$\mathbf{D}_\eta = \begin{bmatrix} \eta_1 & 0 & \cdots & 0 \\ 0 & \eta_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \eta_p \end{bmatrix}, \qquad \mathbf{D}_\alpha = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \alpha_p \end{bmatrix}.$$

Now we will consider the multivariate analogue of the single process law given by

$$\mathbf{H}' = \begin{bmatrix} \mathbf{H}'_\xi \\ \mathbf{H}'_\eta \end{bmatrix} = \begin{bmatrix} (\mathbf{D}_\xi - \mathbf{D}_\mu)^{-1}\ \mathbf{\Delta}'_\xi(\mathbf{D}_\eta - \mathbf{D}_\alpha) \\ (\mathbf{D}_\eta - \mathbf{D}_\alpha)^{-1}\ \mathbf{\Delta}'_\eta(\mathbf{D}_\eta - \mathbf{D}_\alpha) \end{bmatrix}. \tag{42}$$

The analogy in form to the univariate differential equation (13) is evident. A solution of (42) is

$$\begin{aligned} \eta_1 &= \alpha_1 + \beta_1(\xi_a - \mu_a)^{\delta_{1a}}(\xi_b - \mu_b)^{\delta_{1b}} \cdots \\ &\qquad \cdot(\xi_q - \mu_q)^{\delta_{1q}}(1)(\eta_2 - \alpha_2)^{\delta_{12}} \cdots (\eta_p - \alpha_p)^{\delta_{1p}} \\ \eta_2 &= \alpha_2 + \beta_2(\xi_a - \mu_a)^{\delta_{2a}}(\xi_b - \mu_b)^{\delta_{2b}} \cdots \\ &\qquad \cdot(\xi_q - \mu_q)^{\delta_{2q}}(\eta_1 - \alpha_1)^{\delta_{21}}(1) \cdots (\eta_p - \alpha_p)^{\delta_{2p}} \\ \cdots &\quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ \eta_p &= \alpha_p + \beta_p(\xi_a - \mu_a)^{\delta_{pa}}(\xi_b - \mu_b)^{\delta_{pb}} \cdots \\ &\qquad \cdot(\xi_q - \mu_q)^{\delta_{pq}}(\eta_1 - \alpha_1)^{\delta_{p1}}(\eta_2 - \alpha_2)^{\delta_{p2}} \cdots (1). \end{aligned} \tag{43}$$

Thus (43) are the "structural equations" for the causal network arising from allowing each pathway to individually follow the single process law.

As with linear path regression analysis many of the pathways may be non-existent for any particular problem. These restrictions on the structural equations are made simply by setting the corresponding δ to zero.

Note that the structural equations (43) relate "true" variables and parameters. No stochastic variables have been considered yet, and thus, we may transform the equations arbitrarily. We take logs as follows:

$$\begin{aligned} \eta_1^* &= \log(\eta_1 - \alpha_1), & \xi_a^* &= \log(\xi_a - \mu_a), & \alpha_1^* &= \log \beta_1, \\ \eta_2^* &= \log(\eta_2 - \alpha_2), & \xi_b^* &= \log(\xi_b - \mu_b), & \alpha_2^* &= \log \beta_2, \\ \cdots &\quad \cdots & \cdots &\quad \cdots & \cdots &\quad \cdots \\ \eta_p^* &= \log(\eta_p - \alpha_p), & \xi_q^* &= \log(\xi_q - \mu_q), & \alpha_p^* &= \log \beta_p, \end{aligned}$$

and $\alpha_{ij}^* = \delta_{ij}$. Then (43) becomes

$$\begin{array}{l} \eta_1^* = \alpha_1^* + \alpha_{1a}^*\xi_a^* + \alpha_{1b}^*\xi_b^* + \cdots + \alpha_{1q}^*\xi_q^* + \alpha_{12}^*\eta_2^* + \cdots + \alpha_{1p}^*\eta_p^* \\ \eta_2^* = \alpha_2^* + \alpha_{2a}^*\xi_a^* + \alpha_{2b}^*\xi_b^* + \cdots + \alpha_{2q}^*\xi_q^* + \alpha_{21}^*\eta_1^* + \cdots + \alpha_{2p}^*\eta_p^* \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ \eta_p^* = \alpha_p^* + \alpha_{pa}^*\xi_a^* + \alpha_{pb}^*\xi_b^* + \cdots + \alpha_{pq}^*\xi_q^* + \alpha_{p1}^*\eta_1^* + \alpha_{p2}^*\eta_2^* + \cdots \end{array} \qquad (44)$$

These transformed structural equations are identical in form with those underlying linear path regression analysis. For a more detailed comparison see Turner [1959]. We may write (44) in matrix notation as follows:

$$(\mathbf{A}_1\ ,\ \mathbf{A}_2)\begin{bmatrix} \boldsymbol{\xi} \\ \mathbf{n} \end{bmatrix} = \mathbf{n} \qquad (45)$$

where

$$(\mathbf{A}_1\ ,\ \mathbf{A}_2) = \left[\begin{array}{cccc|cccc} \alpha_1^* & \alpha_{1a}^* & \cdots & \alpha_{1q}^* & 0 & \alpha_{12}^* & \cdots & \alpha_{1p}^* \\ \alpha_2^* & \alpha_{2a}^* & \cdots & \alpha_{2q}^* & \alpha_{21}^* & 0 & \cdots & \alpha_{2p}^* \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_p^* & \alpha_{pa}^* & \cdots & \alpha_{pq}^* & \alpha_{p1}^* & \alpha_{p2}^* & \cdots & 0 \end{array}\right],$$

and

$$\boldsymbol{\xi}' = (1\ \xi_a^* \cdots \xi_q^*),$$

$$\mathbf{n}' = (\eta_1^*\ \eta_2^* \cdots \eta_p^*).$$

From (45) we get by multiplication

$$\mathbf{A}_1\boldsymbol{\xi} + \mathbf{A}_2\mathbf{n} = \mathbf{n} \qquad (46)$$

and

$$\mathbf{n} - \mathbf{A}_2\mathbf{n} = \mathbf{A}_1\boldsymbol{\xi}$$

$$(\mathbf{I} - \mathbf{A}_2)\mathbf{n} = \mathbf{A}_1\boldsymbol{\xi}.$$

If $(\mathbf{I} - \mathbf{A}_2)$ is non-singular, we get the reduced transformed structural equations.

$$\mathbf{n} = (\mathbf{I} - \mathbf{A}_2)^{-1}\mathbf{A}_1\boldsymbol{\xi}. \qquad (47)$$

Thus, the transformed secondary variables $\mathbf{n}$ are expressed as linear functions of the transformed primary variables $\boldsymbol{\xi}$ and it is seen that the necessary and sufficient condition for this to be possible is that $(\mathbf{I} - \mathbf{A}_2)$ be non-singular.

Inverse log transformation will then provide reduced structural equations in terms of the original variables also such that each secondary variable is a function of just the primary variables.

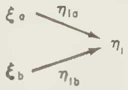
There are several types of causal relationships which are conveniently distinquished. First of all there is the situation in which a single secondary variable is determined by one or more primary variables. This is the case of "ordinary regression", the single structural equation being the non-linear analogue of that corresponding to the usual multiple regression model. A second case is the situation in which two or more secondary variables are jointly determined by some or all of a common set of primary variables. We term this case the case of "joint" regression. In a third situation there is a chain of cause and effect leading from one or more primary variables to a secondary variable and then on to still another "secondary" variable. This case we term the case of "chain" regression. A final type involves cycles of causation or "feedback" from one secondary variable to another and back again after possibly passing through one or more other secondary variables. This is the case of "cyclic" regression. All of these cases individually or in combination have been synoptically treated in what has gone before. Here we wish only to note that the several cases are distinguished by having different kinds of $(\mathbf{I} - \mathbf{A}_2)$ matrices. For this reason we term $(\mathbf{I} - \mathbf{A}_2)$ the "classification" matrix. It may be verified that regression type and $(\mathbf{I} - \mathbf{A}_2)$ are related as follows:

regression type	$(\mathbf{I} - \mathbf{A}_2)$
Ordinary	1 (scalar)
Joint	I
Chain	Triangular
Cyclic (feedback)	Non-triangular

The heavy algebra implied by (47) may be conveniently by-passed by the use of algorithms (see Turner and Stevens, [1959]) when the classification matrix $(\mathbf{I} - \mathbf{A}_2)$ is triangular, that is, when there is no feedback in the system. General algorithms are also available.

We now consider three elementary systems to illustrate the foregoing and to provide concreteness.

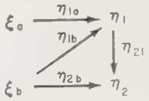
Consider first, the "multiple regression" analogue, with the path diagram below.



The structural equation is then found to be

$$\eta_1 = \alpha_1 + \beta_1(\xi_a - \mu_a)^{\delta_{1a}}(\xi_b - \mu_b)^{\delta_{1b}}. \tag{48}$$

A second example involves ordinary, joint, and chain regression.



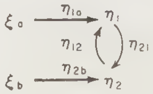
The structural equations are

$$\begin{aligned} \eta_1 &= \alpha_1 + \beta_1(\xi_a - \mu_a)^{\delta_{1a}}(\xi_b - \mu_b)^{\delta_{1b}}, \\ \eta_2 &= \alpha_2 + \beta_2(\xi_b - \mu_b)^{\delta_{2b}}(\eta_1 - \alpha_1)^{\delta_{21}}. \end{aligned} \tag{49}$$

Reduction by use of the algorithms leads to

$$\begin{aligned} \eta_1 &= \alpha_1 + \beta_1(\xi_a - \mu_a)^{\delta_{1a}}(\xi_b - \mu_b)^{\delta_{1b}}, \\ \eta_2 &= \alpha_2 + \beta_2\beta_1^{\delta_{21}}(\xi_a - \mu_a)^{\delta_{21}\delta_{1a}}(\xi_b - \mu_b)^{\delta_{2b}+\delta_{21}\delta_{1b}}. \end{aligned} \tag{50}$$

A final example is one involving feedback. The path diagram is as below.



We have the structural equations

$$\begin{aligned} \eta_1 &= \alpha_1 + \beta_1(\xi_a - \mu_a)^{\delta_{1a}}(\eta_2 - \alpha_2)^{\delta_{12}}, \\ \eta_2 &= \alpha_2 + \beta_2(\xi_b - \mu_b)^{\delta_{2b}}(\eta_1 - \alpha_1)^{\delta_{21}}. \end{aligned} \tag{51}$$

Reduction by (47) then gives

$$\begin{aligned} \eta_1 &= \alpha_1 + \beta_1^{1/(1-\delta_{12}\delta_{21})}\beta_2^{\delta_{12}(1-\delta_{12}\delta_{21})} \\ &\qquad \cdot(\xi_a - \mu_a)^{\delta_{1a}/(1-\delta_{12}\delta_{21})}(\xi_b - \mu_b)^{\delta_{12}\delta_{2b}/(1-\delta_{12}\delta_{21})}, \\ \eta_2 &= \alpha_2 + \beta_1^{\delta_{21}/(1-\delta_{12}\delta_{21})}\beta_2^{1/(1-\delta_{12}\delta_{21})} \\ &\qquad \cdot(\xi_a - \mu_a)^{\delta_{21}\delta_{1a}/(1-\delta_{12}\delta_{21})}(\xi_b - \mu_b)^{\delta_{2b}/(1-\delta_{12}\delta_{21})}. \end{aligned} \tag{52}$$

Let us assume independent additive normal errors in the observable variable y and non-existent errors in the observable variable x so that we have $x_i = \xi_i$ for $i = 1, 2, \cdots, q$, and $y_i = \eta_i + \epsilon_i$ for $i = 1, 2, \cdots, p$. We further assume that the ϵ_i are uncorrelated from one variable to another. Then after reduction we will have p independent non-linear regression equations, each of the form

$$\hat{y} = B_0 + B_1(x_a - M_a)^{D_a}(x_b - M_b)^{D_b} \cdots (x_q - M_q)^{D_q} \tag{53}$$

where D_a, D_b, $\cdots$, D_q are functions of the $\hat{\delta}$'s and B_1 is a function of the $\hat{\beta}$'s and $\hat{\delta}$'s.

Now precisely the same problem of "identification" which plagues linear path analysis arises again here. The problem of estimation is much simplified if we can confine our attention to finding the B's and D's. In certain cases it is then possible to solve for estimates of the original parameters in terms of just the B's and D's. When this is possible, we say the system is *just-identified.* See Turner and Stevens [1959] for discussion. In the present instance we will restrict consideration to the just-identified case.

Again we make use of the Gaussian iterant. We first obtain trial estimates $M_{a0}, M_{b0}, \cdots, M_{q0}, D_{a0}, D_{b0}, \cdots, D_{q0}$. Then we compute:

$$\begin{aligned}
Z_{10} &= (x_a - M_{a0})^{D_{a0}}(x_b - M_{b0})^{D_{b0}} \cdots (x_q - M_{q0})^{D_{q0}},\\
Z_{20} &= Z_{10}/(x_a - M_{a0}),\\
Z_{30} &= Z_{10}/(x_b - M_{b0}),\\
\cdots &\quad \cdots \quad \cdots\\
Z_{q+1,0} &= Z_{10}/(x_q - M_{q0}),\\
Z_{q+2,0} &= Z_{10} \log_e (x_a - M_{a0}),\\
Z_{q+3,0} &= Z_{10} \log_e (x_b - M_{b0}),\\
\cdots &\quad \cdots \quad \cdots\\
Z_{2q+1,0} &= Z_{10} \log_e (x_q - M_{q0}),
\end{aligned} \tag{54}$$

for all observational vectors. In addition we define:

$$\begin{aligned}
B_2 &= -B_{10}\, D_{a0}(M_a - M_{a0}), & B_{q+2} &= B_{10}(D_a - D_{a0}),\\
B_3 &= -B_{10}\, D_{b0}(M_b - M_{b0}), & B_{q+3} &= B_{10}(D_b - D_{b0}),\\
\cdots &\quad \cdots \quad \cdots & \cdots &\quad \cdots \quad \cdots\\
B_{q+1} &= -B_{10}\, D_{q0}(M_q - M_{q0}), & B_{2q+1} &= B_{10}(D_q - D_{q0}).
\end{aligned} \tag{55}$$

Using the device of the total derivative we obtain the approximate linearization

$$\hat{y} \doteqdot B_0 + B_1 Z_{10} + B_2 Z_{20} + \cdots + B_{2q+1} Z_{2q+1} \tag{56}$$

with which to replace (53). By the usual methods of linear regression we now obtain the estimates $B_0, B_1, \cdots, B_{2q+1}$ and from (55) we get the improved estimates $M_a, \cdots, D_q$. These now can be used to replace $M_{a0}, \cdots, D_{q0}$ in (54) and the entire process reiterated until stable estimates are obtained.

If any subset of parameters is known a priori then, as in the simple regression case of Section 3, the appropriate Z's are simply omitted and the analysis carried out without them. For example, if it is known that all processes have vertical asymptotes zero, then we omit Z_{20} through $Z_{q+1,0}$.

When there is more than a single response variable, the application of (56) separately to each provides multiple estimates for the vertical asymptotes M_a, M_b, $\cdots$, M_q. In order to provide unique, efficient estimates of these values, we fit, instead of (56), the family of hyperplanes

$$\hat{y}_i \risingdotseq B_{0i} + B_{1i}Z_{10i} + B_2Z_{20i} + \cdots + B_{q+1}Z_{q+1,0i} + B_{q+2,i}Z_{q+2,0i} + \cdots + B_{2q+1,i}Z_{2q+1,0i} \tag{57}$$

where $i = 1, 2, \cdots, p$ and B_2, B_3, $\cdots$, B_{q+1} are common for all i. Linear regression theory again provides the minimum variance unbiassed estimators for fitting (57).

Now the final estimates of the B's, M's and D's will not be minimum variance unbiassed estimates but they will be maximum likelihood estimates and hence they will be consistent, efficient and invariant under transformation. Thus, since we have assumed a condition of "just-identification" we can solve for estimates of the δ's and these estimates will also be maximum likelihood estimates.

If we were to have an under-identified system, then the methods of this section would be as satisfactory as any. However, if we were to have an over-identified system, another procedure would be required. Perhaps it would be best in that case to simply determine the maximum likelihood estimators directly from the special case being considered. A general computing method appears to be too cumbersome for all possibilities of over-identification.

Asymptotic tests and confidence limits for the just-identified case are found as before by applying linear results to the final iteration. The parabolic test is also applicable to the general case of this section, but, of course, the numerical difficulties are much increased. These may not be prohibitive in the simplest cases, however.

REFERENCES

Deming, W. E. [1943]. *The Statistical Adjustment of Data.* John Wiley and Sons, Inc. New York.

Fisher, R. A. [1956]. *Statistical Methods and Scientific Inference.* Hafner Publishing Co. New York.

Hotelling, H. [1939]. Tubes and spheres in n-spaces and a class of statistical problems. *Amer. J. Mathematics 61*, 440–60.

Pimentel – Gomes, F. [1953]. The use of Mitcherlich's regression law in the analysis of experiments with fertilizers. *Biometrics 9*, 498–516.

Stevens, W. L. [1951]. Asymptotic regression. *Biometrics 7*, 247–67.

Tukey, J. W. [1954]. Causation, regression, and path analysis. *Statistics and*

Mathematics in Biology (O. Kempthorne *et al.*, Eds.). The Iowa State College Press. Ames, Iowa.

Turner, M. E. [1959]. *The single process law: a study in nonlinear regression.* Dissertation No. 59-6571. University Microfilms, Inc. Ann Arbor, Mich.

Turner, M. E., and Stevens, C. D. [1959]. The regression analysis of causal paths. *Biometrics 15*, 236–58.

Wegstein, J. H. [1958]. Accelerating convergence of iterative processes. *Communications of the Assoc. for Computing Machinery 1*, 9–13.

Whitaker, E., and Robinson, G. [1944]. *The Calculus of Observations*, 4*th* edit. Blackie and Son, Ltd. London.

Will, H. S. [1936]. On a general solution for the parameters of any function with application to the theory of organic growth. *Ann. Math. Stat. 7*, 165–90.

Williams, E. J. [1959]. *Regression Analysis.* John Wiley and Sons, Inc. New York.

Wright, S. [1918]. On the nature of size factors. *Genetics 3*, 367–74.

Wright, S. [1921]. Correlation and causation. *J. Agri. Res. 20*, 557–85.

ON THE ANALYSIS OF SPLIT-PLOT EXPERIMENTS

H. Leon Harter
Aeronautical Research Laboratory,
Wright-Patterson Air Force Base, Ohio, U.S.A.

Summary.

A crucial question in the analysis of split-plot experiments is whether or not the interaction between subplot treatments and replications should be pooled with the three-factor interaction of main-plot treatments, subplot treatments, and replications, the result being called subplot error. A brief history of the disagreement over this question is given, along with a rule for deciding, on the basis of a preliminary test of significance, whether or not to pool. Several numerical examples are cited, and one of these is worked out in detail.

1. *History.*

The simplest type of split-plot experiment is one in which a levels of one treatment (the main plot treatment A) are arranged in a randomized blocks design with r blocks (replications R), and b levels of a second treatment (the subplot treatment B) are assigned at random, one to each of b subplots in each of the ar whole plots. It is assumed that a correlation ρ exists between the experimental errors for any two subplots in the same whole plot, but that experimental errors for subplots in different whole plots are uncorrelated. Historically, most statisticians have assumed that interactions between treatments and replications do not really exist. Hence they have computed a sum of squares for subplot error without showing a breakdown into sums of squares for $B \times R$ and $A \times B \times R$, and have used the mean square for subplot error for testing the significance of both B and $A \times B$. Some (see, for example, Federer [5, p. 274]) have even asserted that $B \times R$ and $A \times B \times R$ are confounded with each other and hence are really not separable, though it is possible to compute the two interactions arithmetically. On the other hand, Anderson and Bancroft [1, p. 350] have opined that if the subplot treatments are considered to be random, it is necessary to separate $B \times R$ and $A \times B \times R$. Among those who have considered the case in which the interactions between treatments and replications are not assumed to be zero are Wilk and Kempthorne

[10, pp. 101–102]. Chew [3, pp. 45–49] has assumed only that the three-factor interaction is zero. In either of these cases, $B \times R$ and $A \times B \times R$ must be separated, since the expected values of their mean squares are not identical. We shall consider cases in which it is not known a priori whether or not the interactions between treatments and replications are zero. We shall assume that R (replications) is always a random factor, and consider four cases: (1) A and B both random; (2) A fixed and B random; (3) A random and B fixed; and (4) A and B both fixed. Of these, (4) seems to occur most often in practice. We shall define the variance of an effect involving a fixed factor having f levels with $(f - 1)$ instead of f in the denominator. The expected mean squares are summarized in Table 1, along with the test ratios for A, B, and $A \times B$.

TABLE 1

SUMMARY OF EXPECTED VALUES OF MEAN SQUARES AND TEST RATIOS FOR SPLIT-PLOT EXPERIMENTS

Source	Mean Square	Expected Value of Mean Square (all factors random)*							
		σ_a^2	σ_r^2	σ_{ar}^2	σ_b^2	σ_{ab}^2	σ_{br}^2	σ_{abr}^2	σ^2
Main-plot									
Treatments A	S_a	br		b		$[r]$		[1]	$1 + \overline{b - 1}\rho$
Replications R	S_r		ab	(b)			$[a]$	[(1)]	$1 + \overline{b - 1}\rho$
$A \times R$	S_{ar}			b				[1]	$1 + \overline{b - 1}\rho$
Subplot									
treatments B	S_b				ar	(r)	a	(1)	$1 - \rho$
$A \times B$	S_{ab}					r		1	$1 - \rho$
$B \times R$	S_{br}						a	(1)	$1 - \rho$
A $\times B \times R$	S_{abr}							1	$1 - \rho$

*For brevity, we show multipliers of components of variance in the table; thus, for example, $E(S_{abr}) = \sigma^2{}_{abr} + (1 - \rho)\sigma^2$. If A is fixed, omit components whose coefficients are enclosed in parentheses. If B is fixed, omit components whose coefficients are enclosed in brackets. (It is assumed that R is always a random factor.)

Test Ratios for A

If B is fixed, use S_a/S_{ar} (exact test). If B is random, use S_a/S_{ar} (assumes $\sigma_{ab}{}^2 = 0$) or S_a/S_{ab} (assumes $\sigma^2{}_{ar} = 0$, $\rho = 0$). Satterthwaite test: $S_a/(S_{ar} + S_{ab} - S_{abr})$; Cochran test: $(S_a + S_{abr})/(S_{ar} + S_{ab})$.

Test Ratios for B

If A is fixed, use S_b/S_{br} (exact test). If A is random, use S_b/S_{br} (assumes $\sigma_{ab}{}^2 = 0$) or S_b/S_{ab} (assumes $\sigma^2{}_{br} = 0$). Satterthwaite test: $S_b/(S_{br} + S_{ab} - S_{abr})$; Cochran test: $(S_b + S_{abr})/(S_{br} + S_{ab})$.

Test Ratio for $A \times B$

Use S_{ab}/S_{abr} (exact test)
(One is usually not interested in testing significance of R and its interactons.)

2. *Decision Rule for Pooling.*

We find in Table 1 that the expected value of the mean square for the $A \times B \times R$ interaction is $\sigma^2_{abr} + (1 - \rho)\sigma^2$. The expected value of the mean square for the $B \times R$ interaction is $a\sigma^2_{br} + \sigma^2_{abr} + (1 - \rho)\sigma^2$ if A is random, but it is $a\sigma^2_{br} + (1 - \rho)\sigma^2$ if A is fixed. In either case, the expected values of the mean squares for $B \times R$ and for $A \times B \times R$ are not identical; hence they should not be pooled unless this action is indicated by the results of a preliminary test of significance. This preliminary test is a test of the null hypothesis $H_0 : E(S_{br}) = E(S_{abr})$. Since R is assumed to be random, both $S_{br}/E(S_{br})$ and $S_{abr}/E(S_{abr})$ have central χ^2/d. f. distributions. Hence under H_0 the test ratio S_{br}/S_{abr} has a central F distribution. The form of this preliminary test depends upon whether A is random or fixed. If A is random, H_0 reduces to $H_0' : \sigma^2_{br} = 0$, the expected value of the mean square for $B \times R$ is greater than or equal to the expected value of the mean square for $A \times B \times R$, and the required preliminary test of significance is a one-sided test on the ratio of the mean square for $B \times R$ to the mean square for $A \times B \times R$. If this ratio does not exceed a certain percentage point of the F distribution, the two mean squares are pooled; otherwise, they are not pooled. This problem has been studied by Bozivich, Bancroft and Hartley [2], who recommend a preliminary test of size α, where $.25 \leq \alpha \leq .50$. On the other hand, if A is fixed, H_0 reduces to $H_0'' : \sigma^2_{abr} = a\sigma^2_{br}$, nothing can be said about the relative magnitude of the expected mean squares for $B \times R$ and $A \times B \times R$, and the procedure recommended by Bozivich, Bancroft and Hartley must be modified slightly; the preliminary test of significance is a two-sided test of size α, with probability $\alpha/2$ in each tail of the F distribution. In this case, it is not possible to make a test of size $\alpha = .25$ without computing a special table, since no table of the upper 12.5 percent points of the F distribution is available. Tests of size $\alpha = .20$ and $\alpha = .50$ are possible, using tables of the upper 10 percent and 25 percent points of the F distribution tabulated by Merrington and Thompson[7].

3. *Examples.*

Five numerical examples have been taken from the literature and the subplot error for each has been subdivided into $B \times R$ and $A \times B \times R$ interactions. In each of these examples, it appears reasonable to assume that both A and B are fixed factors, so that the preliminary test of significance will be two-sided. Table 2 gives, for each example, the source of the data, the mean squares and the numbers of degrees of freedom for $B \times R$ and for $A \times B \times R$, the ratio of the larger mean

TABLE 2

PRELIMINARY TESTS FOR POOLING IN SPLIT-PLOT EXPERIMENTS

Source Code	$B \times R$ Interaction		$A \times B \times R$ Interaction		Ratio = Larger M. S. / Smaller M. S.	Critical Values	
	M. S.	d.f.	M. S.	d.f.		$F_{.10}$	$F_{.25}$
(1)	0.0367	15	0.0236	30	1.56	1.72	1.32
(2)	15.29	6	25.50	42	1.67	2.78	1.75
(3)	28.08	9	13.18	27	2.13	1.87	1.37
(4)	119.21	15	206.02	30	1.73	1.87	1.40
(5)	19.36	70	21.03	140	1.09	1.32	1.16

Sources: (1) Snedecor [9, Table 11.35, p. 310], (2) Federer [5, Table X−2, p. 276], (3) Anderson and Bancroft [1, Table 23.2, p. 348], (4) Rider [8, Exercise 4, p. 191], taken from Yates [11], (5) Cochran and Cox [4, Table 7.5, p. 225].

square to the smaller and the upper 10 percent and 25 percent points of the F distribution. If we use a preliminary test of size $\alpha = .50$, we conclude that $B \times R$ and $A \times B \times R$ should not be pooled for the data from sources (1), (3), and (4). Even if we take $\alpha = .20$, we decide not to pool in the case of the data from source (3). Let us now look at this example in greater detail. The data represent yields (in bushels per acre) of corn in an experiment conducted to compare four methods of seedbed preparation and four methods of planting, using four blocks. The methods of seedbed preparation are the whole-plot treatments A, the planting methods are the subplot treatments B, and the blocks are the replications R. Since there are two fixed factors (A and B)

TABLE 3

ANALYSIS OF VARIANCE FOR DATA FROM SOURCE (3) ABOVE

Source of Variation	d.f.	S.S.	M. S.	F
Seedbed methods (A)	3	194.56	64.85	$64.85/17.58 = 3.69$
Replications (R)	3	223.81	74.60	
$A \times R$	9	158.25	17.58	
Planting methods (B)	3	4107.39	1369.13	$1369.13/28.08 = 48.76$**
$A \times B$	9	221.74	24.64	$24.64/13.18 = 1.87$
$B \times R$	9	252.72	28.08	
$A \times B \times R$	27	355.75	13.18	
Total	63	5514.22	—	

**Significant at the one percent level.

and only one random factor (R), the fixed main effects and interaction are tested by their interactions with the random factor, as shown in Table 3. Anderson and Bancroft give the correct test for the whole-plot treatments A, but test the subplot treatments B and the interaction $A \times B$ by the subplot error obtained by pooling $B \times R$ and $A \times B \times R$. In this case they reach the correct conclusions (B highly significant, $A \times B$ non-significant), but clearly this will not always be true.

4. *Strip Plots and Sub-subplots.*

The expected mean squares are different for strip-plot experiments than for ordinary split-plot experiments, but the same tests are appropriate (see, for example [4], [6]), and the same preliminary test for pooling can be used. The theory can also be extended in an obvious way to split-split-plot experiments. If it cannot be assumed that the whole-plot treatment A, the subplot treatment B, and the sub-subplot treatment C do not interact with replications, the sub-subplot error should be partitioned into four parts $C \times R$, $A \times C \times R$, $B \times C \times R$ and $A \times B \times C \times R$. In the most common case (A, B and C fixed, and R random), these are appropriate for testing C, $A \times C$, $B \times C$ and $A \times B \times C$ respectively. They should be pooled only if this is indicated by the results of appropriate preliminary tests of significance.

5. *Acknowledgments.*

The author gratefully acknowledges the benefits received from discussion of this problem with his colleagues, Paul R. Rider and Mary D. Lum, and from studying the comments of the editor of this Journal and two referees.

REFERENCES

[1] Anderson, R. L. and Bancroft, T. A. [1952]. *Statistical Theory in Research.* McGraw-Hill Book Company, Inc., New York.

[2] Bozivich, Helen, Bancroft, T. A., and Hartley, H. O. [1956]. Power of analysis of variance test procedures for certain incompletely specified models, I. *Ann. Math. Stat.*, *27*, 1017–43.

[3] Chew, Victor. [1958]. *Experimental Designs in Industry.* John Wiley & Sons, Inc., New York.

[4] Cochran, William G. and Cox, Gertrude M. [1950]. *Experimental Designs.* John Wiley & Sons, Inc., New York.

[5] Federer, Walter T. [1955]. *Experimental Design.* The Macmillan Company, New York.

[6] Kempthorne, Oscar. [1952]. *Design and Analysis of Experiments.* John Wiley & Sons, Inc., New York.

[7] Merrington, Maxine and Thompson, Catherine M. [1943]. Tables of percentage

points of the inverted Beta (F) distribution. *Biometrika, 33,* 73–88 (reprinted in part as Table 18 in Pearson, E. S. and Hartley, H. O. [1956]. *Biometrika Tables for Statisticians*, Volume I, University Press, Cambridge.)

[8] Rider, Paul R. [1939]. *An Introduction to Modern Statistical Methods.* John Wiley & Sons, Inc., New York.

[9] Snedecor, George W. [1946]. *Statistical Methods.* Fourth Edition, The Iowa State College Press, Ames.

[10] Wilk, M. B. and Kempthorne, O. [1955]. Analysis of Variance: Preliminary Tests, Pooling, and Linear Models—Derived Linear Models and their Use in the Analysis of Randomized Experiments. Technical Report 55–244, Volume II, Wright Air Development Center.

[11] Yates, F. [1935]. Complex experiments. *Suppl. Jour. Roy. Stat. Soc. 2,* 181–247.

QUERIES AND NOTES

D. J. Finney, *Editor*

156 NOTE: Confidence Intervals for Recombination Experiments with Microorganisms

A. W. Kimball[1]
Mathematics Panel and Biology Division,
Oak Ridge National Laboratory,[2]
Oak Ridge, Tennessee, U.S.A.

1. METHODOLOGY

In a typical recombination experiment, a parent bearing a biochemical mutation at the A locus is mated with a parent bearing a biochemical mutation at the B locus, both loci on the same chromosome. The two parental types are thus $(a+)$ and $(+b)$. The loci (or markers) are usually identified with nutrient elements such as adenine or lysine, and a mutation at one of these loci signifies that the organism carrying the mutation will not grow unless the corresponding nutrient element is present in the growth medium. Progeny from such matings are of four types. If no recombination takes place on the chromosome between the loci, progeny must be of the parental types, and the two classes have equal expected frequencies. If a single recombination takes place in this region, the progeny will be $(++)$ or (ab) and these types likewise have equal expected frequencies. Types $(++)$ and (ab) are called recombinants and types $(a+)$ and $(+b)$ are called non-recombinants. Clearly the frequency of occurrence of recombinants is dependent on the distance between the loci, assuming that recombination is equally likely to occur anywhere along the chromosome, and obtaining information about distances between loci is one of the main purposes of a recombination experiment. The proportion of recombinants is called the recombination frequency and the proportion of the $(++)$ type is called the prototroph frequency.

Since, ordinarily, the four types of progeny cannot be separately identified, platings are made both on complete media (containing elements A and B) and on minimal media (A and B omitted). All four

[1] Present address, Department of Biostatistics, The Johns Hopkins University.

[2] Operated by Union Carbide Corporation for the U. S. Atomic Energy Commission.

types will grow on complete media but only the prototroph (++) will grow on minimal media. Thus, if a count of x organisms is found by plating v_x ml of a suspension on a minimal medium and a count of y organisms by plating v_y ml on a complete medium, the prototroph frequency is estimated by

$$R = v_y x / v_x y.$$

If the total count, $S = x + y$, is large and neither x nor y is too small, a confidence interval can be computed from the formulas,

$$R_L = v_y(T - A)/v_x(1 + A), \qquad R_U = v_y(T + A)/v_x(1 - A) \tag{1}$$

where R_L and R_U, respectively, are the lower and upper confidence limits, $T = x/y$, $A = t_\alpha(T/S)^{1/2}$, and where t_α is the normal deviate corresponding to a $100(1 - \alpha)$ percent confidence interval.

As a numerical example, consider an experiment that yielded the values $v_x = 180$, $x = 561$, $v_y = 2.5$, $y = 3759$ for which 95 percent confidence limits ($t_{.05} = 1.96$) are required. We have

$$R = \frac{(2.5)(561)}{(180)(3759)} = .002073$$

$$T = 561/3759 = .14924, \qquad S = 561 + 3759 = 4320,$$

$$A = 1.96(.14924/4320)^{1/2} = .01152,$$

$$R_L = \frac{(2.5)(.14924 - .01152)}{180(1 + .01152)} = .001891,$$

$$R_U = \frac{2.5(.14924 + .01152)}{180(1 - .01152)} = .002259.$$

These limits are not, in general, symmetrical about the estimate.

In many cases, the experimenter will have preliminary information about the expected orders of magnitude of the ratios x/v_x and y/v_y and may want to know how large an experiment is required to yield a reasonably precise estimate of the prototroph frequency. One measure of the precision is the relative width of the confidence interval,

$$W = (R_U - R_L)/R.$$

Ordinarily, one is interested in relative widths of not more than 30 percent ($W \leq .30$), and in this case the relation,

$$S = 4t_\alpha^2(1 + T)^2/TW^2, \tag{2}$$

holds approximately. Both the total count S and the ratio of the counts T, are important in determining the precision of the estimate, and it can be shown that if the counts are in a 1 : 1 ratio, the total count required for a specified relative interval width will be a minimum. With this in mind, if we put $T = 1$ in (2), the total count required is given by

$$S = 16t_\alpha^2/W^2. \tag{3}$$

Thus the procedure is to decide on the confidence level and interval width desired and to calculate the total count from (3). Then, knowing that the counts should be in a 1 : 1 ratio and having preliminary information about x/v_x and y/v_y, the amounts to be plated out can be calculated directly. If the amounts calculated turn out to be experimentally unfeasible, one might have to depart from a 1 : 1 ratio for the counts, in which case, the total count can be calculated from (2) for whatever value of T is found to be acceptable. Some trial and error with the use of (2) may be necessary in these cases.

As an example, consider a situation in which it is expected that the counts from a standard suspension will be about 2 per *ml* for x and 5,000 per ml for y. A relative width of about 20 percent is desired for a 95 percent confidence interval. Thus $t_\alpha = 1.96$, $W = .20$ and

$$S = 16(1.96)^2/(.20)^2 = 1540,$$

whereupon the amounts to be plated should be chosen so that approximately 770 counts will be obtained from the platings on each medium. In other words, $v_x = 770/2 = 385$ and $v_y = 770/5{,}000 = 0.15$ are the amounts to be plated on minimal and complete media, respectively.

Although sample sizes determined by this method may be used as a rough guide in designing recombination experiments, they should not be regarded as optimum. Readers interested in statistically more sophisticated procedures for determining sample sizes should refer to Birnbaum [1954].

2. DERIVATION

Since v_x and v_y are known constants, the statistical problem consists of finding a confidence interval for the parameter $\theta = \mu/\nu$, where $\mu = E(x)$ and $\nu = E(y)$. Assuming that x and y are independently distributed Poisson variables, Birnbaum [1954] showed that a confidence interval for θ may be constructed by treating $p = y/(x + y)$ as a binomial variable with $E(p \mid x + y) = \pi = \nu/\mu + \nu$. Limits on π, say (π_0, π_1), are easily obtainable from published tables of the binomial distribution, whereupon limits on θ are given by

$$\theta_0 = (1/\pi_1) - 1 \quad \text{and} \quad \theta_1 = (1/\pi_0) - 1.$$

In many cases, however, $x + y$ is large and the normal approximation to the binomial distribution may be used to calculate the limits on π, providing π is not near zero or one. This in turn leads to the limits given in (1).

The relative width of the confidence interval for the prototroph frequency can be written as

$$W = W_\pi(1 + T)/T[1 - (W_\pi^2/4)]$$

where $W_\pi = (\pi_1 - \pi_0)/p$, the relative interval width for π. In most practical situations W_π will be small, say $\leq .30$, in which case $W \cong W_\pi[1 + (1/T)]$. When the substitution is made for W_π, the result can be expressed as in (2). For fixed W, this expression has a minimum at $T = 1$, and (3) follows directly. I am indebted to the referee for pointing out that if $W \leq .30$, W_π will be about half as large as W when $T = 1$, and the approximation will be even better than indicated.

REFERENCE

Birnbaum, A. [1954]. Statistical methods for Poisson processes and exponential populations. *J. Amer. Stat. Assoc. 49*, 254–66.

157 NOTE: Estimation of Proportion from Zero-truncated Binomial Data

G. N. Wilkinson

Division of Mathematical Statistics,
Commonwealth Scientific and Industrial Research Organization,
University of Adelaide, South Australia

INTRODUCTION

An important problem of estimation arises, chiefly in genetical contexts, when the proportion of individuals bearing some particular character must be estimated from the frequencies observed in a random sample of groups or families which have at least one individual bearing this character. The estimation by maximum likelihood of the proportion p in this situation has been discussed principally by Haldane [5], [6], Fisher [3] and Finney [2].

In conjunction with D. G. Catcheside's studies on tetraploid maize [1] the author derived an iterative procedure, which is presented here, with a simpler mathematical form and rather better convergence properties than the standard method based on score and information, which is described by Finney [2].[1]

It was found preferable to estimate, instead of p, a parameter M related to p by the equation $X/M = p$, where X is the total number of character-bearers in the sample. The parameter M may be thought of as the total number of individuals that might have been observed if families with no character-bearers had also been ascertained in the sample. In this respect the method described here is a special case of the general method given by Hartley [7], in which the application of maximum likelihood to data with missing (or coalesced) classes is facilitated by estimation of the relevant frequencies for those classes. A similar approach has also been used by Fyfe [4].

The formula given below, however, leads to less computation than Hartley's procedure, when several different sizes of family are involved. In the present instance accelerated convergence has been obtained by applying Newton's method of approach by tangents to the solution of the equation $M = M^*(M)$, and requires, for each cycle of computation, only a simple iteration in this equation. Hartley, on the other hand, employs the "approach to the geometric limit", which differs from Newton's method in that approximating chords are used in place of tangents, and thus requires *two* basic iterations in each cycle.

THE ITERATION FORMULA

The basic sampling distribution is the zero-truncated binomial. Accordingly, the likelihood function for the data, using the notation given in Table 1, is essentially

$$L = X \log p + (N - X) \log q - \sum_{s \geq 2} f_s \log (1 - q^s). \qquad (1)$$

If $M = X/p$, $m_s^* = n_s/(1 - q^s)$, and $M^* = \sum m_s^*$, the differential coefficient of the likelihood function with respect to M, may be expressed in the form

$$\frac{dL}{dM} = -\frac{p}{Mq}(M - M^*). \qquad (2)$$

[1] In Finney's paper it is suggested, with reference to the tables of scoring functions provided, that linear interpolation may be used to obtain more accurate approximations to the maximum likelihood value $\hat{p}$. As a matter of principle it may be noted that *linear* interpolation is insufficient for this purpose. Linear interpolation, between the entries for p_1 and p_2 say, will determine an estimate intermediate to the two estimates based on p_1 and p_2 respectively, but generally the latter estimates will both lie on the same side of the maximum likelihood value $\hat{p}$, even when p_1 and p_2 are themselves on opposite sides of $\hat{p}$. More accurate interpolation, based on second differences as well, is therefore necessary.

The factor p/Mq is essentially non-zero. Application of Newton's method, in which the successive approximations to the zero of $f(x)$ are given by the formula

$$x_{k+1} = x_k - [f(x_k)/f'(x_k)],$$

to the factor $(M - M^*)$ gives the formula

$$M_{k+1} = (M_k^* - r_k M_k)/(1 - r_k), \tag{3}$$

where

$$r = \frac{p}{Mq} \sum_{s \geq 2} \frac{m_s^*(m_s^* - n_s)}{f_s}.$$

TABLE 1

NOTATION

Size of group, s	Frequency	Number of individuals	Number of character-bearers
2	f_2	$n_2 = 2f_2$	x_2
3	f_3	$n_3 = 3f_3$	x_3
. . .	. . .	. . .	. . .
s	f_s	$n_s = sf_s$	x_s
. . .	. . .	. . .	. . .
Total	F	N	X

Estimation is completed by determining $\hat{p}$ as the proportion of X to $\hat{M}$, and the amount of information in relation to p may then be determined as

$$I_p \doteq \frac{M}{pq}(1 - r), \tag{4}$$

evaluated at the point of maximum likelihood. The amount of information differs from that for simple binomial data by the factor $(1 - r)$. The factor r thus emerges as the *relative* amount of information lost by truncation in the sample. This factor governs the rate of convergence of the iterative process, convergence being extremely rapid when r is small. Even when the loss of information is as high as 40 percent, however, as in the example below, a single cycle of computation, from a suitable starting point, will give a high degree of accuracy.

A very good starting point is provided by the (iterative) solution of the much simpler equation

$$M = N/(1 - q^{\bar{s}}), \tag{5}$$

where $\bar{s}$ is the mean size of families observed, or the equivalent equation for the Poisson distribution (if p is small), as used by Fyfe [4]. Indeed for many practical purposes the solution of (5) will itself be sufficiently accurate. In the example given below, in spite of the wide range of family sizes, the solution of (5), roughly $M = 3{,}300$, is within 3 percent of the correct maximum likelihood value, and, in the case of the data on human albinism discussed by Haldane [6], the relative error is 4 percent.

EXAMPLE

Catcheside [1] gives the numbers of recessive *su* seeds observed on 24 cobs from tetraploid maize triplex for su—1^+ and pollinated by nulliplex. It was possible that the parent plants of one or more of the seven cobs observed with no *su* seeds might have been quadruplex

TABLE 2

Iterative Solution of the Equation $M = M^*(M)$

$s, = n_s$	x	q_0^s	m^*_{s0}	q_1^s	m^*_{s1}
18	1	.7935	87.2	.83749	110.76
65	2	.4338	114.8	.52707	137.44
74	1	.3864	120.6	.48235	142.95
76	1	.3766	121.9	.47294	144.20
84	2	.3399	127.2	.43709	149.22
86	2	.3312	128.6	.42856	150.50
131	2	.1858	160.9	.27508	180.71
140	3	.1655	167.8	.25174	187.10
173	2	.1083	194.0	.18187	211.46
181	2	.0977	200.6	.16808	217.57
188	2	.0893	206.4	.15688	222.98
206	1	.0709	221.7	.13139	237.16
211	1	.0665	226.0	.12507	241.16
219	2	.0600	233.0	.11559	247.62
223	4	.0570	236.5	.11112	250.88
251	3	.0398	261.4	.08433	274.12
259	2	.0359	268.6	.07794	280.89
$N = 2585$	$X = 33$	$M_0^* = 3077.2$		$M_1^* = 3386.72$	
$S = \sum m_s^* (m_s^* - n_s)/f_s$ (all $f_s = 1$, $F = 17$) $M_0 = 2585$ $p_0 = 0.012766$		$S_0 = 73{,}842$ $r_0 = 0.3694$ $1 - r_0 = 0.6306$ $M_1 = 3{,}366$ $p_1 = 0.009804$		$S_1 = 142{,}276$ $r_1 = 0.41850$ $1 - r_1 = 0.58150$ $M_2 = 3{,}401.63$ $p_2 = 0.009701$	

rather than triplex. As a check, therefore, p was estimated from the 17 cobs with one or more *su* seeds. The iterative calculations for these data are set out in Table 2, and give an estimate $\hat{M} = 3402$ in very close agreement with the actual number of seeds observed, 3359, on all 24 cobs. The sensitivity of the data on this point is rather low, however, as the relative standard error of $\hat{M}$ (and $\hat{p}$), determined by formula (4), is 23 percent.

To exhibit the convergence properties, and for comparative purposes, $M_0 = N = 2585$ has been chosen as the starting point for the computations in Table 2. In practice the value 3359 might have been chosen, or, lacking this information, the approximate solution 3300 of the simpler equation (5). A single iteration would then have given far more accuracy than actually required. The values of q^s were determined in this instance by logarithmic calculation, but this calculation can be avoided if suitable binomial tables are available. The values r_k are determined essentially by the computation of a weighted sum of products of the m_s^* with the differences $m_s^* - n_s$ (not shown).

To compare rates of convergence the successive approximations to the solution of $dL/dp = (M - M^*)/q$, by adjustment in terms of score and information, were also determined. The comparison is shown in Table 3. It can be seen that the present method does very much better from a poor start than the other, requiring only one iteration instead of two. However the disparity diminishes, as one would expect, in the neighbourhood of the maximum likelihood value.

TABLE 3

SUCCESSIVE APPROXIMATIONS TO $\hat{M} = 3401.76$ BY APPLICATION OF NEWTON'S METHOD TO THE EQUATIONS (i) $M = M^*(M)$, (ii) $dL/dp = 0$. RELATIVE ERRORS ARE SHOWN IN PARENTHESES

	M_0	M_1	M_2	M_3
$M = M^*$	2585	3366 (1%)	3401.63 (0.004%)	3401.76 ($\hat{M}$)
$dL/dp = 0$	2585	3697 (10%)	3421.19 (0.6%)	3402.04 (0.008%)

COMMENT

The approach which led to the present method was somewhat different from Hartley's and may be of interest in other connections. In the simplest cases the differential of the likelihood function with respect to a parameter (or some related parameter), θ say, can be factorized into two parts, one linear in θ and the other essentially non-zero. More generally,

$$dL/d\theta = g(l(\theta, x)).h(\theta, x)$$

where $l(\theta, x)$ is linear in θ, $g(0) = 0$, and $h(\theta, x)$ is essentially nonzero. In such cases one has the explicit solution given by the zero of $l(\theta, x)$ and no iteration is necessary. It was argued that, in the less simple cases, a similar factorization might be sought, in which $l(\theta, x)$, though not strictly linear, was in some sense asymptotically linear. In the present instance, for example, the function $M - M^*$ is asymptotically linear in M as $\bar{x} = X/F$ increases. A sketch graph of this function is shown in Figure 1. Curvature of the function is restricted by a tangent and an asymptote which differ in slope only by the factor $1/\bar{x}$. Generally, it could be expected that iterative methods applied to $l(\theta, x)$ rather than to $dL/d\theta$ would give more rapid convergence, as the functions g and h in the latter will generally have an adverse effect on convergence by increasing curvature. The comparison is complicated, however, by the presence of one or more inflexion points in $dL/d\theta$.

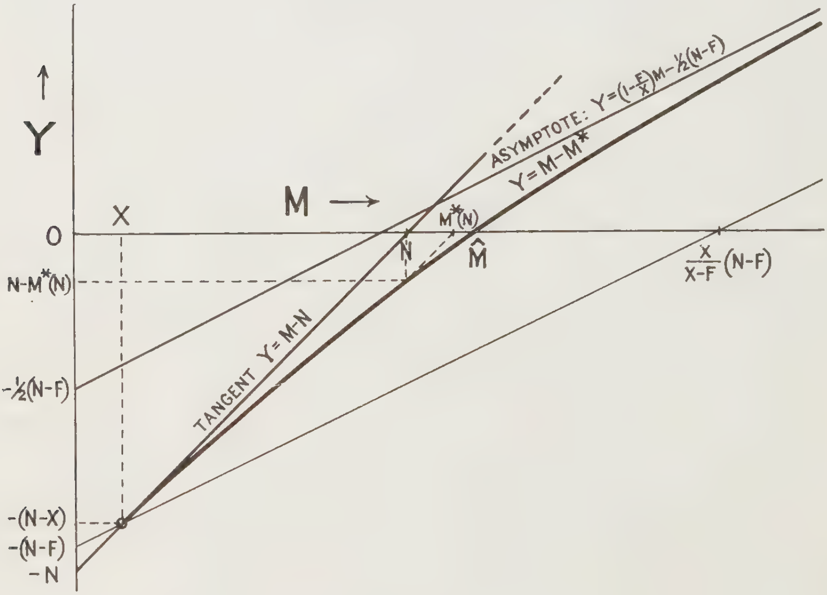


FIGURE 1

Sketch Graph of the Function $M - M^*(M)$

(The intercept $X(N - F)/(X - F)$ corresponds to Haldane's estimate of p, [6], from data in which the probability of ascertaining a family is proportional to the number of affected individuals in the family.)

REFERENCES

[1] Catcheside, D. G. [1956]. Double reduction and numerical non-disjunction in tetraploid maize. *Heredity 10*, 205.

[2] Finney, D. J. [1949]. The truncated binomial distribution. *Ann. Eugen. 14*, 319.

[3] Fisher, R. A. [1936]. The effects of methods of ascertainment upon the estimation of frequencies. *Ann. Eugen. 6*, 13.

[4] Fyfe, V. C. [1953]. Double reduction at the Mid locus in Lythrum salicaria. *Heredity 7*, 285.

[5] Haldane, J. B. S. [1932]. A method for investigating recessive characters in man. *J. Genet. 25*, 251.

[6] Haldane, J. B. S. [1938]. The estimation of the frequencies of recessive conditions in man. *Ann. Eugen. 8*, 255.

[7] Hartley, H. O. [1958]. Maximum likelihood estimation from incomplete data. *Biometrics 14*, 174.

BOOK REVIEWS

J. G. Skellam, *Editor*

Members are Invited to Suggest Books for Listing or Review to the Editor

7 GUMBEL, E. J. **Statistics of Extremes.** New York: Columbia University Press, 1958. Pp. xx + 375. $15.00.

B. F. Kimball, *New York Public Service Commission, New York, N. Y., U.S.A.*

This work is essentially a source book on the statistics of extreme values. Furthermore it is the only source which is comprehensive. Viewed as a source book it is well arranged and well documented. The bibliography includes over 600 books and articles, to which specific references are made throughout the book. Of these, 34 are papers by the author, or written in collaboration.

The first chapter deals with definitions and the basic concepts. In the second chapter, on *order statistics*, the theory of *exceedances* is set forth and handled in a straight-forward and concise manner.

In Chapter 3 the exact distribution of the extreme values of a sample are introduced. Chapter 4 is given over to an analysis of the exact distributions of extremes, pointing out the essentially different types which are important.

These first four chapters comprise nearly half the subject matter and include all the material dealing with "exact" distributions of extreme sample values. Perhaps the principal advances in the study of the distribution of extremes in the last twenty five years have been brought about by recognition that the asymptotic forms of these distributions are usually adequate as approximations to the exact distributions in the applied field, and are much simpler to deal with.

The asymptotic forms fall naturally into three classes, of which the doubly exponential is of the most consequence since it is the asymptotic form which results from a universe having an "exponential" type of cumulative distribution function, such as the logistic, normal and Gamma distributions. A second class, or "second asymptote" as referred to by Gumbel, stems from distributions of the Cauchy or Pareto type. The third class is often referred to as Weibull's distribution when used for the lower extreme. Both can be obtained from the first by suitable logarithmic transformations. The first "asymptote" and its applications are discussed in Chapters 5 and 6, covering about 100 pages. This class is of particular interest to biologists—for example, in studying extinction times of bacteria[1]. The second and third asymptotes are treated in Chapter 7. Chapter 8 is devoted to the asymptotic forms of the range and mid-range, and their applications.

[1] *Statistical Theory of Extreme Values and Some Practical Applications*, U. S. Dept. of Commerce N. B. S. Applied Math. Series 33(1954) p. 43; U. S. Gov. Printing Office, Washington, D. C. 40c.

This reviewer finds the author very ingenious in deriving methods which he believes will facilitate application of the theory to practical problems. Some of these ingenious devices are open to criticism. For example, the discussion of the topic *Fitting Straight Lines on Probability Papers* (pp. 34–36) is more ingenious than logical.

A formula is developed at the bottom of page 23 for the probability that a return period T will be reached in v trials. This probability is "distribution free" and for large T can be approximated by $W(v) = 1 - \exp(-v/T)$. From this Gumbel develops what he calls a "control interval".

$T/\lambda < v < T\lambda$, $\lambda > 0$, the probability that v will lie on this interval being given by $P = e^{-1/\lambda} - e^{-\lambda}$. This is ingenious, but, since the probability density function is monotonically decreasing from $v = 0$, the shortest interval carrying the same probability would be $0 < v < T\lambda'$ with $P = 1 - e^{-\lambda'}$. Hence, for $P = .6827$, $\lambda' = 1.148$ and the "control interval" would be $0 < v < 1.148T$ rather than $(.3196)T < v < (3.129)T$.

A similar criticism applies to the control intervals (5) on page 215. These intervals are centered at the mode. The "shortest" such interval for $P = .6827$ would be given by $\Delta = -.917$, $\Delta = +1.317$, rather than by $\Delta = \pm 1.141$. This would change (8) to $T(y + \Delta) = 3.732T(y) - 1.366$, $T(y - \Delta) = 0.400T(y) + 0.300$. Likewise for $P = .9545$, $\Delta' = \pm 3.0069$ becomes $\Delta' = -1.59$ and 3.251. Thus, when the underlying distribution is asymmetrical, some adjustment should be made when possible.

In presenting *The Extreme Control Band* of Table 6.1.6, p. 218, and in other references to the control band, besides noting the lopsided character of the "shortest interval" of fixed probability, the author might have pointed out that most of these formulae presume a *correct* estimate of the scale parameter α. As I pointed out in my contribution to the book, the uncertainty of the estimate of α will involve the variance as a function of y, the distance from the mode, and N, the size of the sample used in determining α [cf. (2) p. 235]. Thus the complete variance upon which a confidence band should be based will involve both the "variance of position" of the fitted line and the intrinsic variance of the order statistic (used as a basis for Table 6.1.6.)[2].

Other formulas for the "variance of position" are casually referred to, such as the "Control Curve of Dick and Darwin", Section 6.1.7; and formula (5), p. 228, which gives the sampling variance of an estimate x obtained by the method of moments. In using this latter formula it should be made clear that the σ^2 in the denominator refers to the variance of the sampled values and hence might be replaced by s^2. The formula (6) which follows it would appear to be in error. With $(\pi^2/6)/\alpha^2$ substituted for σ^2 and $\sigma^2(y) = \alpha^2\sigma^2(x)$ the α^2 should cancel out. Also the significance of (6) is not clear since y is usually *assigned* a value in making an estimate from (4) by the method of moments.

The matter of the identification of the variance of the order statistic vis à vis that of the sampling variance of the estimate (variance of position) needs to be made clearer.

I have noticed only two rather obvious typographical errors: On p. 24, line 6, .4637 should be replaced by .04657, and on p. 215, line 1, .05450 should be replaced by .95450.

[2] Henry Schultz, The standard error of a forecast from a curve. *J. Amer. Stat. Asso. 25* (1930), 139–85.

VESSEREAU, A. **Méthodes Statistiques en Biologie et en Agronomie,** Tome
8 deuxième. Paris: J.-B. Baillière et Fils, Editeurs. Nouvelle edition entièrement
refondue, 1960. Pp. 540. 55 NF.

P. Robinson, *Canada Department of Agriculture, Ottawa, Ontario, Canada.*

This book, as the title implies, is intended principally for research workers in the biological sciences. There are now several good texts available in English, but there are few in the French language. In his introduction to the first edition the author says "Le présent ouvrage se propose de combler cette lacune." That there was this need may perhaps be judged by the short time elapsing between the first and second editions. The second edition has been completely revised and now includes—

(i) a fuller discussion of the underlying mathematical reasoning,
(ii) a more general description of the analysis of variance.
(iii) a more extensive treatment of experimental design, including confounding, and
(iv) tables and graphs relating to the binomial and Poisson distributions.

This book contains a great deal of useful information and the mathematical treatment is kept simple throughout. Basic ideas on probability and sampling are discussed in the first four chapters, simple tests of significance in the next two chapters, experimental design in the next nine, and regression and correlation in the last two.

In covering such a wide field within the confines of a single book, any treatment in depth is of course precluded. The broad approach to the subject may have led the author to include descriptions of higher order moments, the β_1 and β_2 measures of skewness and kurtosis, and Sheppard's corrections. These, however, are completely unnecessary in an introductory text of this nature intended for the biologist, as is evidenced by the fact that they do not reappear after Chapter II. It is also a pity that outdated measures of dispersion—the probable error (pp. 56 and 100) and the mean error (p. 100)—are discussed. These are interesting from a historical viewpoint, but only serve to confuse the agricultural worker.

It would have been better to sacrifice breadth of approach in these areas in order to deal a little more fully with problems of more direct concern to the biologist. In particular the sections on transformations, split plots, and the combined analyses of repeated experiments could have been expanded. A discussion on expected mean squares might have helped when dealing with this last problem, the correct choice of the appropriate error for testing a particular hypothesis would then have been clearer, instead of appearing to be left to the whim of the experimenter (p. 419).

A fuller discussion of the difficulties involved in multiple comparisons would also have been helpful. Comments (2) and (3) on pp. 187 and 188, for example, appear to be contradictory, and the discussions on pp. 203, 204, 206 and 216 could be misleading.

The use of artificial data in certain illustrations (pp. 193, 458) is unwarranted when so many examples are now available. This is particularly dangerous when dealing with the analysis of covariance because the interpretation can be tricky. The author's suggestion of circumstances in which a covariance analysis could be used—p. 451—"Lorsqu' on étudie l'influence des engrais sur l'indice de pureté du sucre extrait d'une culture de betterave sucrière, les résultats dépendent du poids des racines, qui n'est pas constant dans les différentes parcelles de l'essai, mais qui peut être mesuré après coup"—is a particular case in point.

Finally I must disagree with a piece of advice which the author gives in his "En guise de conclusion." He suggests, if there is any doubt in the mind of the research worker as to the applicability of some of the techniques discussed, that "mieux vant encore s'en tenir aux schémas éprouvés." This value of this advice depends entirely upon what is meant by "schémas éprouvés". In general, it would be far better to advise the research worker "de demander conseil aupres des experts en statistiques."

9 LOVE, A. G., HAMILTON, E. L. and HELLMAN, I. L. **Tabulating Equipment and Army Medical Statistics**, Washington, D. C.: Office of the Surgeon-General, U. S. Army Medical Service, 1958. Pp. x + 202. 36 Figures. $2.00.

W. W. Holland, *London School of Hygiene and Tropical Medicine, London, England*

This is a delightfully written account of the development of machine methods and of the techniques of recording in the American Army. The book not only traces the development of statistical services, but also shows how great the dependence of an efficient military medical service is on the collection, tabulation and assessment of data. After reading this short monograph it is easy to understand how the excellent description of such diseases as cholera, respiratory infection and of injuries sustained by American Army personnel during the two World Wars could be completed. It is of interest to read that, even in the first years of the existence of the American Army, during the Napoleonic Wars, acute respiratory disease was a severe problem. Although the treatment tends, at times, to be somewhat too biographical, a good account of the development of punch cards, sorting and tabulating systems and of computers is given. Full details and illustrations of the various forms used for recording data are described as well as an account of the collection and analysis of anthropometric information obtained during routine assessment of personnel on enlistment and discharge.

ABSTRACTS

The following are abstracts of papers presented at the annual meeting of the Biometric Society ENAR, held jointly with the Biometric Society WNAR, the American Statistical Association, and the Institute of Mathematical Statistics in Stanford, California on August 23 to 26, 1960. Abstracts of additional papers presented at this meeting appear in other society publications.

683 KLAUS ABT (U. S. Naval Weapons Laboratory, Dahlgren, Va.). **Analysis of Variance of Differences Versus Analysis of Covariance.**

In analyzing paired data $(x; y)$, classified in one or more directions, where x and y are observations of the same variable (and therefore usually observed at different times) at least two methods are available. First there is the analysis of covariance in which y is considered to be stochastically dependent on x, the "independent" variable. A second method is the analysis of variance of the generalized differences, $d = y - \beta_0 x$, briefly called "analysis of differences". The case $\beta_0 = 1$ represents the most important case of ordinary differences.

It is shown that, if the condition $E(b_E) = \beta_0$ is satisfied ("expectation of the coefficient of error sum of squares regression line equals β_0", which easily can be tested), the analysis of differences has a power equal to or even higher than that of the analysis of covariance. From many calculated examples, it seems that the condition mentioned is satisfied quite often. In many cases, therefore, the analysis of differences will be the appropriate method for the data to be analyzed, besides this including a minimum of computation.

684 R. L. ANDERSON (North Carolina State College, Raleigh, N. C.). **Recent Developments in Multivariate Analysis.**

One of the most important recent developments in multivariate analysis has been the publication of some excellent textbooks on the subject, on both theoretical and applied aspects. This paper is in the nature of a discussion of what the writer feels is needed to improve multivariate techniques. The chief need is much more attention being paid to construction of experimental models, usually involving two or more equations, and methods of estimating parameters in these models, assessing their usefulness and testing hypotheses concerning them. Particular attention is paid to causal chains and to dynamic systems. For the latter, methods based on lagged dependency models and Laplace transform techniques are indicated.

Some possible weaknesses of current multivariate analysis of variance procedures are indicated. Problems which should be studied in the following fields are outlined: discriminatory, canonical, factor and component analysis. Special attention is focused on the construction of selection indexes, and the selection of predictors in ordinary multiple regression analyses.

685 ROLF E. BARGMANN (Virginia Polytechnic Institute, Blacksburg, Va.). **On the Problem of Ordering Variables in Tracing Significant Contributions.**

After rejection of some general linear hypothesis in multivariate analysis it is frequently desirable to ascertain how much each variable contributes to the observed differences. The step-down procedure—i.e., a process analogous to analysis of covariance in which we study the first variable, the second given the first, the third given the first and second, etc.—depends on the ordering of the variables. The present report discusses the use for this purpose of the correlations of a linear discriminant function $a'X$, where a' is the eigenvector associated with the largest root of HE^{-1} (H = matrix of S.S. and S.P. due to hypothesis, E same due to error). A similar procedure is described for the discrimination of common factors, i.e., for those cases where any difference between groups is supposed to be in terms of those artificial components only which are shared by two or more of the observable variables. A detailed demonstration study will be presented for illustration.

686 EDWARD C. BRYANT and DONALD W. KING (University of Wyoming, Laramie, Wy.). **Estimation from Populations Identified by Overlapping Sampling Frames.**

Consider a population of unknown size identified by p sampling lists, each with $\tilde{M}_i$ entires from which independent samples of size m_i , $i = 1, 2, \cdots , p$ may be drawn. It is assumed that the sampling procedure will reveal which of the p lists contain the individuals interviewed. An estimate of the total number in the population is desired, as well as the number of individuals appearing on each possible combination of sampling lists. Maximum likelihood and minimum modified chi square estimates are considered. By moderate restrictions, an approximation to the variance is provided by use of the Taylor's series expansion around the expected numbers in each category of listing combinations.

687 BRADLEY E. COPELAND (New England Deaconess Hospital, Boston, Mass.). **Problem in Pathology With Statistical Aspects.**

The pathologist in his role in diagnostic laboratory medicine finds a compelling need for a broader comprehension of the fundamental principles of statistics for himself and on the part of the practicing physician, the medical technologist, and the medical research worker.

Every physician must be equipped with certain statistical tools to properly utilize the laboratory measurements now available to him for diagnosis and treatment.

There must be clear and rapid communication between the pathologist and the clinician concerning the variation in normal values and the reliability and inherent variability of laboratory measurements.

The first major problem relates to the barrier of statistical nomenclature and the multiplicity of statistical formulas. Communication cannot be successful until pathologists and clinicians learn to handle with ease such fundamental statistical concepts as standard deviation, standard error, variance, the t test, and the F ratio.

Since medicine is concerned with the individual, there is a need to clearly identify the tools which separate the diseased individual from a normal population as opposed to the tools which differentiate normal and abnormal populations. Both are useful devices, but they are the source of much confusion in the minds of physicians.

Currently there is an also pressing need for more comprehensive definitions of normal values in all laboratory measurements—normal as regards variation in age, sex, geographical location, and any other environmental or hereditary factor which may influence the normal population.

To achieve permanent indoctrination of the medical profession with useful statistical tools is a necessary and useful goal. It is in this area that active communication between pathologists and statisticians should take place to develop solutions to the problems described.

688 R. C. ELSTON (University of North Carolina, Chapel Hill, N. C.). **On Additivity in the Analysis of Variance.**

The problems of testing for non-additivity and of finding the best transformation of the data to minimize it are discussed. Consider the model: $E(y_q) = \mu + \mu_q + c\mu_q^2$, where y_q is an observation belonging to the q-th subclass in an analysis of variance problem. If this model fits our data, and if max μq − min $\mu q < 1/|c|$ (the range over which this model is a strictly monotonic function of μ_q), then there exists a useful transformation of our data that will give additivity; that is the transformed data will be fitted by an additive model. Two tests that have already been proposed for testing for non-additivity are tests of the null hypothesis $c = 0$ against $c \neq 0$ under models that can be considered as approximations to the above model. It is shown how estimates of μ and c can be used as guides in finding an appropriate transformation of the data.

689 WALTER T. FEDERER (Cornell University, Ithaca, N. Y.). **Augmented Designs with Two-, Three- and Higher-Way Elimination of Heterogeneity.**

Augmented experimental designs are standard ones to which additional treatments have been added. The additional treatments may or may not be replicated the same number of times as the treatments in the standard design or the same number of times as other new treatments. The groupings of treatments into rows, columns, complete blocks, etc. may be such that the size of the group is or is not equal. Construction procedures, randomization procedures and the analyses in general form have been presented for augmented experimental designs with two-, three- and higher-way elimination of heterogeneity.

690 WILLIAM R. GAFFEY (California State Department of Public Health). **Tests of Hypotheses Concerning Boundedness in Convolutions.**

Suppose the random variable Y, with unknown d.f. $G(y)$, is observed subject to error, so that $X = Y + Z$ is the random variable actually observed. Let $H_y(z)$ be the (known) conditional d.f. of the continuous "error" random variable Z, and let $F(X)$ be the d.f. of X. We are concerned with deriving tests of the following two-hypotheses about the distribution of the "true" random variable Y; (1) $G(y) = 0$ for $y < a$, and (2) $G(y) = 1$ for $y > b$.

Under suitable restrictions on the form of $H_y(z)$, it turns out that hypothesis (1) is true if and only if $F(X) \leq H_a(x - a)$ for $x < a$, and hypothesis (2) is true if and only if $F(X) \geq H_b(x - b)$ for $x > b$. A test of either inequality is therefore a test of the corresponding hypothesis about $G(y)$.

A test is proposed using the one sided Kolmogorov-Smirnov statistic, restricted to the appropriate portion of the range of X. The large sample distribution of the statistic is derived, and the consistency of the test is established.

691 MAX HALPERIN (General Electric Co., Schenectady, N. Y.). **Nearly Least Squares Estimates in Heteroscedastic Regression.**

Suppose we have observations y_{ij}, $i = 1, \cdots, k$, $j = 1, \cdots, n_i$ where $Ey_{ij} = a + bx_i$ and Var $y_{ij} = \sigma_i^2$, the x_i are presumed known, and the y_{ij} are normally distributed, mutually independent, and $k > 2$, min $n_i > k$. In this paper we show how to generate estimates of a and b which are nearly least squares in the sense that the correlation of the estimates is identical to that which would be obtained from a least squares fit with known variances while the variance of the estimates differs from that appropriate to the known variance least squares fit by a multiplicative factor which approaches unity as $\min_i n_i \to \infty$ and k remains constant. The procedure further allows estimation of the covariance matrix of the estimates with $(\min_i n_i - k + 1)$ degrees of freedom and hence an exact confidence region for a and b based on Hotelling's T^2 distribution. The results cited are of a conditional nature in which the conditioning constrains the variance multiplying factor referred to above. The distribution of the multiplicative factor aside from known constants is that of Hotelling's central T^2 in general (Student's "t" if $k = 3$). As a consequence, alternative exact confidence intervals of unconditioned nature are available in principle by averaging the conditional results over the appropriate distribution. It is conjectured that matching moments to determine approximate degrees of freedom for an "equivalent" T^2 distribution is adequate from a practical point of view. If $k = 2$, the procedure we propose gives estimates with exactly least squares variances and exact confidence intervals or regions. It is conjectured that results similar to the above generalize to general polynomial or polyvariable regression linear in the parameters and a sketch of the type of argument is given.

692 R. C. HENNEMUTH (Inter-American Tropical Tuna Commission). **Estimating Vital Statistics of Yellowfin Tuna Populations.**

Knowledge of vital statistics is essential to the study of the dynamics of exploited fish populations; age composition, growth rate and mortality rates being particularly important.

Determination of age of yellowfin tuna by using marks on scales or bones has not proven reliable; however, growth rate and relative age have been estimated by observing the temporal changes in the size composition of catch. Absolute age has been approximately determined by comparing the average time of spawning with the time and size at which the age groups initially appear in the catch.

Total mortality rate has been estimated by analyzing catch-curves of successive year classes, which have been computed from age composition and indices of apparent abundance based on catch per unit effort data.

693 BURTON J. HOYLE and GEORGE A. BAKER (University of California, Davis, Cal.). **Game Theory Applied to Field Trials.**

The determination of the proper variety of a cereal to grow may be considered as a game with the Grower as one opponent and Nature as the other. A game matrix of forty-one independent yield trials on nine strains of Hannchen barley is presented. The trials cover three years and very diverse designs and environments within the years. A similar but more restricted matrix for kernel weight is also presented. The whole structure of field trials is made clear by a contemplation of these game matrices. Sometimes one strain does better and sometimes another. That is, it is doubtful that there is a best strain. Sometimes all strains are close together and sometimes the strains differ widely. Some strains are very erratic, in relative yield, and some are very consistent. Strains highest with respect to yield are not necessarily good with respect to kernel weight and other desirable properties so over-all ratings involve weighting considerations on various characteristics.

It is clear from dominance considerations that some strains should seldom or never be grown. In other cases mixtures of strains may be better than any one strain.

A determination of the odds for the different strains depends on the distribution of nature's strategies.

694 AUGUSTUS C. JOHNSON (Booz-Allen Applied Research, Inc.). **A Stochastic Model of Incubation-Period Distributions.**

An independent-action birth-death model of infection is presented, based on the hypothesis that each organism has probability $\lambda\, dt$ of dividing and $\mu\, dt$ of dying during the time interval dt. λ and μ being constants. Incubation terminates when a colony reaches a lower absorbing barrier at o or an upper absorbing barrier at some number N, at which point a sign or symptom appears.

The resulting differential-difference system is solved to give expressions for the distribution of conditional probability that n organisms will grow to an upper absorbing barrier N between times o and t. Expressions for the moments are derived both from the distribution and directly, being exhibited explicitly for moments through the third.

605 CECIL L. KALLER and VIRGIL L. ANDERSON (Purdue University, Lafayette, Ind.). **An Environmental Extension of Chromosome Analysis in Population Genetics.**

A phase model has been developed for the special case of a diploid diallelic genetic population having K pairs of chromosomes in genetic equilibrium which are influenced by Q tri-level environmental factors. The theoretical structure for this development is based on the theory of factorial experimental designs with each of the $K + Q$ factors having three "levels." Estimates of the parameters of the population described by this model are obtained by extending the methods presented in a thesis by McKean (Purdue, 1958). This sampling method involves the formulation of a "chromosome population" by the random selection of a set of K chromosome pairs from the original genetic population. The chromosome pairs so drawn are combined to form all possible genotypes; individuals of each resulting genotype are placed in all 3^Q possible environmental level combinations in numbers proportional to their theoretical frequency. Under the assumption that all interactions involving three or more factors are zero, an analysis scheme is developed for the estimation of chromosome population parameters. Therefrom, by the relationship demonstrated between the chromosome population parameters and the corresponding parameters of the original genetic population, conditions are established under which these latter are estimable.

696 MARVIN A. KASTENBAUM (Oak Ridge National Laboratories, Oak Ridge, Tenn.). **Countercurrent Dialysis—A Stochastic Process.**

Peptides isolated from rat liver microsomes may be separated from contaminating amino acids, sugars and salts by countercurrent dialysis. A single such dialysis can achieve only fractional purification. Therefore, more complex systems of multiple dialysis, in which a linear series of cells are used, have been devised. The system of immediate interest may be described as follows: Let each dialysis cell be a stage, and each dialysis period be a cycle. Then after each cycle, the dialout (the solution outside the dialysis sac), at each stage is concentrated and used as the dialin (the solution inside the dialysis sac) in the succeeding stage, whereas the dialin at each stage is concentrated and returned to the dialysis sac of the previous stage. The dialin at the first stage is retained at the first stage, and the dialout at the final stage is taken out of the system. In the nomenclature of stochastic processes, the first stage represents a reflecting barrier and the last stage an absorbing barrier.

To determine the probability with which a particle of the isolate will be at a specified stage in the system after a given number of cycles have been carried out, the system is described in terms of a Markov process with a stochastic matrix resulting from the product of two matrices, namely, a diffusion matrix and a transfer matrix, and algebraically explicit solutions are derived for any number of stages and cycles.

697 JOSEPH KEILIN (U. S. Dept. of Health, Education, and Welfare, Washington, D. C.). **The Use of the Information Statistic as a Measure of Conformity in Comparing Two Sets of Responses.**

A morbidity survey of the city of Nashville and suburbs was recently conducted by the Air Pollution Medical Program. In order to obtain some measure of the reliability of responses a subsample was chosen for reinterview.

Two methods of looking at the data are discussed and two functions are proposed for measuring the proportion of information transmitted from the ordinal interview. The two methods dictate different null hypotheses and these in turn lead to the two different functions.

The results from the reinterview are presented and show relatively low amounts of information transmission. Further analysis of the data, using information theory procedures, is suggested.

698 JOHN G. KEMENY (Dartmouth College, Hanover, N. H.). **General Remarks on Markov Chains with Illustrated Examples in Biology.**

Part I of the paper deals with finite Markov chains. It is shown how, through intelligent classification of various types of chains and the use of modern matrix techniques, many fundamental problems can be reduced to routine computations. Two fundamental matrices are developed, one for absorbing and one for ergodic chains, in terms of which basic quantities can be simply expressed. As applications of these mathematical techniques, problems in genetics are considered.

Part II of the paper deals with some recent joint research carried on with J. L. Snell, concerning denumerable Markov chains. It is shown that extensions of the above-mentioned matrix techniques to denumerable chains result in an interesting discrete analogue of classical potential theory. While the discrete version is simpler than the classical potential theory, it can be vastly generalized by extending it to a wide class of denumerable Markov chains. This has been done previously for transient chains, but the extension to recurrent chains is new. The fundamental matrices extend to basic potential operators. Applications of these results are sketched to population problems and to the spread of epidemics.

Throughout the paper, stress is laid on the methodological advantages of setting up a Markov chain model. A number of key results are stated without proof and several unsolved problems are mentioned.

699 OSCAR KEMPTHORNE (Iowa State University, Ames, Iowa), and R. N. CURNOW (University of Aberdeen, Aberdeen, Scotland). **The Partial Diallel Cross.**

The diallel cross, which is composed of all possible single crosses among a group of inbred lines, is a common plan of investigation in plant and animal breeding. It can be used to estimate general and specific combining abilities and variances, to estimate components of genotypic variation and to estimate yielding capacities of multi-way crosses. With n lines it does however involve $n(n - 1)/2$ crosses, even if maternal effects can be ignored. Breeding programs can easily lead to 50 inbred lines which merit examination, so the diallel cross becomes unfeasible. This paper deals with sampling of the complete diallel cross table in a specific say, namely:

(1) arrange the lines in random order
(2) obtain the crosses of line i with lines

$$k + i, k + i + 1, \cdots, k + i - 1 + s$$

where $k = (n + 1 - s)/2$ is an interger and all number above n are reduced by a multiple of n so as to be between 1 and n.

Each line is crossed then with s lines and $s = n - 1$ corresponds to the complete diallel cross. This partial diallel cross is considered in the following respects:

(1) estimation of variance components
(2) comparing the yielding capacities of all possible crosses and
(3) estimation of general combining abilities.

The efficiency of the plan with regard to each of these aspects is evaluated and discussed. It is shown to be more efficient than some other possible plans under some circumstances.

700 BERTRAM S. KRAUS (University of Washington, Seattle, Wash.). **The Study of Human Growth: Some New Horizons.**

Researchers in biometrics, genetics, and physical anthropology have a fertile field for miscegenation in the study of growth. An example is the timely and significant problem of the possible genetic effects of irradiation on future human generations. Thus far workers in this field have failed to demonstrate that radiation causes mutations in the human species, although this can be assumed from experiments with other mammals. Perhaps this failure can be associated with insufficient knowledge or recognition of the nature of growth and the role of the genetic constitution in controlling or regulating growth. Perhaps, too, there has been too much emphasis placed upon clinical classifications of abnormalities (congential malformations) and not enough on the less dramatic but none the less real deviations in patterns and rates of growth. Equally significant is the failure to define abnormality in terms of the statistical characteristics of the given population.

Attention is focused in this paper upon the distribution of certain skeletal phenomena in prenatal life—sequence of appearance of centers of ossification in the foot, changing long bone inter-relationships, and appearance of centers of calcification in the primary dentition. Norms for human populations are estimated for these traits. It is then suggested that significant deviations from norms established in control populations may reflect higher frequencies of mutant genes in a test population. A method is proposed for the study of the genetic effects of irradiation in Hiroshima and Nagasaki wherein the biometrist, the geneticist, and the physical anthropologist may pool their skills and concepts.

701 B. KURKJIAN (Diamond Ordnance Fuse Laboratory) and M. ZELEN (National Bureau of Standards, Washington, D. C.). **A General Theory for Analysis of Asymmetrical, Confounded Factorial Experiments.**

A general theory is developed for the analysis of factorial experiments involving an arbitrary number of factors and levels of each. The treatment-combinations are assumed to be assigned to an appropriate incomplete block design and the treatment estimates, $\hat{\tau}$'s, are computed in the usual way, ignoring the factorial correspondence between the treatments. New results are presented for obtaining the best linear estimates (and the associated var-cov matrix) of the various main effect and interaction parameters in terms of the $\hat{\tau}$'s. A condition on the var-cov matrix of treatment estimates, $V_{\hat{\tau}}$, is provided which insures that the sum of squares of the various main effect and interaction terms is distributed as chi-square. Furthermore, a certain large class of partially balanced incomplete block (PHIB) designs with arbitary number of associate classes is described which can be treated by methods of this paper. The theory is applied to two PHIB designs involving three associate classes.

702 ROBERT S. LEDLEY (George Washington University, Washington, D. C.). **A Sequential Decision Theory Applied to Medical Diagnosis.**

The reasoning foundations of medical diagnosis involve two well-known mathematical disciplines; symbolic logic and probability theory. Value theory then can aid the choice of an optimum treatment.

Three factors are involved in the logical analysis: medical knowledge relating disease and symptom complexes; the symptom complex of the patient; and the disease complexes that are the final diagnosis. Each of these can be expressed as a Boolean function. The ease of making diagnostic tests varies greatly; a realistic method of diagnosis uses easier tests to select among more difficult tests.

Using probability theory, a 'most likely' diagnosis is determined by applying Bayes's formula to the conditional probability that a disease complex will produce a particular symptom complex and the total probability that any person chosen from the sample under consideration will have that disease complex. Certain problems arise in computing statistics for the probability determinations; the difficulty in obtaining 'sufficiently large' data and the time delay in gathering data.

Value concepts in medical diagnosis and treatment are concerned, for example, with decisions on whether to continue further testing, or what are the values of particular alternative treatment—diagnosis combinations. We distinguish three kinds of problems in treatment decision; under certainty, i.e. when the disease is definitely known; under risk, i.e. when alternative diagnoses have been made with known probabilities; under uncertainty, i.e. when alternative diagnoses remain with no information concerning the probabilities.

703 K. H. LU (Utah State University, Logan, Utah). **The Means and Variances of the Products of Two or Three Normal Variables.**

Given normal variables t_1, t_2 and t_3 with means μ_1, μ_2 and μ_3, variances σ_{11}, σ_{22} and σ_{33} respectively. Let t_1 and t_2 follow the bivariate normal distribution with moment generating function:

$$M(t_1 t_2) = \text{exp.} \sum_{i=1}^{2} \mu_i t_i + \tfrac{1}{2} \sum_{ij=1}^{2} \sigma_{ij} \mu_i \mu_j$$

Also let t_1, t_2, and t_3 follow the trivariate normal distribution with moment generating function:

$$M(t_1 t_2 t_3) = \text{exp.} \sum_{i=1}^{3} \mu_i t_i + \tfrac{1}{2} \sum_{ij=1}^{3} \sigma_{ij} \mu_i \mu_j$$

We define new variables x and y as follows:

$$x = t_1 t_2 , \quad \text{and} \quad y = t_1 t_2 t_3 \quad \text{respectively.}$$

It can be shown that by differentiating $M(t_1 t_2)$ and $M(t_1 t_2 t_3)$ to obtain the appropriate partial derivatives evaluated at $t_i = 0$ and substituting into the definitions of means and variances, we have:

$$\mu_x = \mu_1 \mu_2 + \sigma_{12} ,$$

$$\sigma_{xx} = \sigma_{11} \mu_2^2 + \sigma_{22} \mu_1^2 + 2 \mu_1 \mu_2 \sigma_{12} + \sigma_{11} \sigma_{22} (1 + \rho_{12}^2),$$

$$\mu_y = \mu_1 \mu_2 \mu_3 + \mu_1 \sigma_{23} + \mu_2 \sigma_{13} + \mu_3 \sigma_{12} ,$$

$$\begin{aligned} \sigma_{yy} = {} & \sigma_{11} \mu_2^2 \mu_3^2 + \sigma_{22} \mu_1^2 \mu_3^2 + \sigma_{33} \mu_1^2 \mu_2^2 + \sigma_{11} \sigma_{22} \mu_3^2 (1 + \rho_{12}^2) \\ & + \sigma_{11} \sigma_{33} \mu_2^2 (1 + \rho_{13}^2) + \sigma_{22} \sigma_{33} \mu_1^2 (1 + \rho_{23}^2) \\ & + \sigma_{11} \sigma_{22} \sigma_{33} (1 + 2\rho_{12}^2 + 2\rho_{13}^2 + 2\rho_{23}^2 + 8 \rho_{12} \rho_{13} \rho_{23}) \\ & + 2(\sigma_{12} \mu_1 \mu_2 \mu_3^2 + \sigma_{13} \mu_1 \mu^2 \mu_3 + \sigma_{23} \mu_1^2 \mu_2 \mu_3) \\ & + \mu_1 \mu_2 (6 \sigma_{23} \sigma_{13} + 4 \sigma_{12} \sigma_{33}) + \mu_1 \mu_3 (6 \sigma_{12} \sigma_{23} + 4 \sigma_{13} \sigma_{22}) \\ & + \mu_2 \mu_3 (6 \sigma_{12} \sigma_{13} + 4 \sigma_{23} \sigma_{11}). \end{aligned}$$

704 H. M. C. LUYKX and BETTY L. MURRAY (office of the Surgeon General U.S.A.F.). **Durations of Illness.**

Using epidemiology in the broad sense of the word, one of the epidemiological characteristics of a disease is its duration pattern. Not only the mean duration will vary from one diagnostic entity to the next, but so will the shape of the distribution of durations, which can be described in several ways. In this paper a picture is presented for each of a number of specific diseases. For example, graphs are shown for 368 cases of infectious hepatitis, 446 cases of mumps, 879 cases of pilonidal cyst and 2177 cases of streptococcal sore throat. These and other charts illustrate the techniques of analysis and presentation, and the use made of the data.

The curves are "cases—remaining" curves, which are strictly analogous to survivorship curves commonly given with life tables. A cohort begins its episode of illness at the zero point on a scale of days, and diminishes as each case terminates. The cases—remaining curves are more useful than ordinary frequency polygons (on a scale of duration). They show at a glance the range and skewness of the duration distribution, the modal day for return to duty, and the median and other percentile points on the "days-after-admission" scale.

The Air Force medical data are unique in that they provide unusually large frequencies, derived from uniform records and recording procedures, showing diagnoses established by a physician in each case. In this presentation, the beginning of an illness is excusal from duty by a physician, and the end is return to military duty; cases not returned to duty (deaths or disability separations from the service) are not included.

Definitions of illness, diagnostic entity, disability, admission, return to duty, etc., are also discussed, and a technical note describes the method of computing confidence intervals for percentile points.

705 CLIFFORD J. MALONEY (U. S. Army Chemical Corps, Ft. Dietrick, Md.). **Disease Severity Quantitation, II.**

It was previously shown (Disease Severity Quantitation. Paper read at Washington, D. C. 1959 Annual Meeting of the American Statistical Association) that symptoms of rats exposed to aerosol challenge of *B tularense* conform closely to a scaling model due to Guttman (Stouffer, S. A., Measurement and Prediction, 1950, Princeton U. Press).

Necessary and sufficient criteria of scalability are that symptom pairs yield four-fold contingency tables in which one, but not both, of the secondary diagonal cells are zero, and that the zero cells in all tables can be arranged in a linear order in either symptom designator.

The present paper points out that (in the absence of experimental accidents) the sum of the four cells in every table is fixed, that each table divides the entire disease severity scale into three intervals. In the first, the disease severity is too slight for either symptom to appear. In the second, the more sensitive symptom appears, but not the other, and in the last segment, the disease severity is such that both symptoms appear. The lower boundary is fixed by the milder symptom and is located at the same severity level whatever the other member of the pair; similarly for the upper level. Hence, a system of linear equations can be formed in the segment frequencies and in the segment lengths.

These sets can be equated and expressed in matrix notation. If the experimental data is perfect, the matrix will be triangular. If the data are subject to random error, the further treatment can be carried out in terms of a matrix subject to random error.

706 H. E. MCKEAN and B. B. BOHREN (Population Genetics Institute, Purdue University, Lafayette, Indiana. U. S. A.). **Numerical Aspects of the Regression of Parent on Offspring.**

The relative merits of three methods of estimating heritability from the regression of offspring on one parent's record, from the point of view of statistical efficiency, are brought out with respect to five generations of a closed poultry flock previously reported by Yamada, Bohren, and Crittenden (*Poultry Science*, 1957). The three methods discussed are: (1) The regression of offspring means on parent records. (2) The regression of offspring records on parent records, in which the parents' record is repeated once for each progeny. (3) The Kempthorne—Tandon weighted technique (*Biometrics*, 1953), in which account is taken of the correlation coefficient ρ between errors associated with progeny of the same parent.

The latter technique (3) depends upon knowing ρ exactly; in this case the estimation technique is optimum. If ρ is unknown, ρ must be estimated or guessed; in which case (3) is not optimum. The loss in efficiency in misguessing ρ is discussed, over a range of possible values of ρ. A comparison is simultaneously made with methods (1) and (2).

The general conclusions, based on these data are: (i) method 1 is distinctly inferior to methods 2 and 3 in most conceivable situations; (ii) if ρ is small (which it is if the genetic variance is primarily additive and maternal effects are minor), then method (3) is only slightly more efficient than method (2). Indeed, method (2) is simply a special case of method (3) with ρ always guessed to be 0.

707 HARLEY B. MESSINGER (University of California, Berkeley, Cal.). **A Geometrical Model for Height-Weight Control Charts in Schoolchildren.**

The primary objective of this study was to obtain a satisfactory screening device, i.e., a way to give a yes-or-no answer at any time to the question, "Are this child's height and weight within the ranges one would expect knowing his age, sex, and previous heights and weights$_{y}$" To accomplish this, a set of tolerance limits were derived from the combination of cross-sectional studies of two functions of height and weight in one healthy group of children and longitudinal studies of these functions in each individual of a second healthy group with respect to the cross-sectional standards set by the first group. Because the functions were approximately independent, separate control charts could be used.

A secondary objective of the study was to seek interpretability for the control charts. In an interpretable chart, a child's ratings at any time can be related directly to his physical characteristics. A class of geometrically-meaningful functions arising from the use of a cylinder to approximate the human body was examined in the Institute of Child Welfare "Guidance Study" data. The height (H) of the cylinder and the diameter of the cylinder relative to the height (D/H) were essentially uncorrelated within each age group from 4 to 14 years. Individual children of the "Berkeley Growth Study" were studied longitudinally. Their H and D/Y functions at each age were studentized to obtain standard- or z-scores. From studies of these sequences as stationary autoregressive time series, reasonably efficient control charts were obtained. The D/H function, called "relative broadness," can be gotten from a nomogram. No arithmetic computations or standard-score tables are needed for this kind of chart.

708 CHARLES J. MODE (Montana State College, Bozeman, Mont.). **On The Theory of the Improvement of Metric Traits In Outbreeding Populations.**

An attempt was made in this paper to characterize the change in the mean and genetic variance of a metric trait under a program of improvement. The approach used in this paper differs from other approaches by carefully distinguishing between the fitness of a genotype under a program of improvement and the metric trait. Only the single locus case with an arbitrary number of alleles was considered, but the model may be interpreted in such a way as to include the multi-locus case.

Steady state populations and the problem of stability of a steady state were also considered. The existence of steady states and the question of their stability was discussed for constant and non-constant measure of fitness.

The conditions for the existence of an optimal value for the mean of a metric trait when two or more alleles are maintained in a population under a program of improvement were also given.

709 W. S. OVERTON and A. L. FINKNER (North Carolina State College, Raleigh, N. C.). **The Sample Road-Block Method of Estimating Hunting Pressure.**

The sampling method described was developed for use on Game Management Areas in Florida, but is applicable in situations where estimates of hunting pressure are desired for areas having few or no inhabitants. The method involves the sampling of access points.

Under a given definition, it is possible to express "hunting pressure" in terms of total persons entering and leaving the area by time period. An expression is developed for the definition used in Florida, which forms the structural basis for sampling. Total hunting pressure is divided into a daily component and a component due to camping; only the former is considered in this paper.

The daily component of hunting pressure is a simple linear combination of the time period totals of entries and exits. There are a number of sampling systems that will give unbiased estimates of the totals by time periods, and, hence, unbiased estimates of the daily component. A system in which an observer measures a randomly selected station at alternate time periods with the starting time randomized among days was chosen because of its simplicity and ease of operation. Time periods are considered strata, and the usual stratified sample estimates of totals and variances are employed. This system does not yield unbiased estimates of variance, but seems to be satisfactory for present purposes.

Alternative sampling systems and corresponding estimates of variance are considered and evaluated.

710 D. S. ROBSON (Cornell University, Ithaca, N. Y.). **Cumulant component analysis in balanced designs.**

Cumulant components, as a direct extension of variance components, described the manner in which the usual assumptions on the linear model are violated. The unique, minimum variance unbaised estimators of cumulant components for the distribution-free, balanced model are easily determined because of the existence of a complete, sufficient order statistic. One computational procedure for constructing these estimates is an algebraic extension of the analysis of variance to the "analysis of cumulants".

Under the usual Model II assumptions of normality and independence of effects, the components of all cumulants beyond the second are zero. Tests of departure from zero, therefore, constitute tests of the usual assumptions. Several such tests are available.

711 B. V. SHAH (Iowa State University, Ames, Iowa). **Asymmetrical Factorials in Incomplete Blocks.**

The primary purpose of this paper is to summarize the recent developments in this subject and to indicate possible lines of attack for future research. After a brief summary on history of developments, the desirable properties of factorial experiments, namely orthogonality and balance are discussed. The analysis and the construction of the designs with above properties are studied. As a consequence of relaxing the condition on balance we obtain a wider class of designs, which may be called partially balanced factorial experiments. Finally, the use of pseudo-factors and some other aspects of these designs are studied.

712 R. G. D. STEEL (Cornell University, Ithaca, N. Y.). **A Variable Rank Sum Multiple Comparisons Test.**

A variable rank sum is used for the comparison of all pairs of treatments. This sequential method of testing should be more powerful than a procedure using a fixed rank sum and an experimentwise error rate. Computation of exact probabilities and appropriate tables is discussed.

713 PATRICK SUPPES (Stanford University, Stanford, Cal.). **Foundations of Measurement.**

This paper is concerned with the axiomatic foundations of measurement. The procedures and operations of measurement can be represented as a relation algebra in the sense of Tarski. The two formal theorems which demonstrate the adequacy of a particular scheme of measurement are the representation theorem which establishes that the algebra of empirical relations can be mapped homomorphically into the appropriate structure of the real numbers, and the uniqueness theorem which establishes that the homorphism is unique up to the appropriate group of transformations. Some empirical examples of these ideas are given and some problems of axiomatizability are mentioned.

714 WILLIAM F. TAYLOR (University of California, Berkeley, Cal.) and JOSEPH BERKSON (Mayo Clinic, Rochester, Minn.). **A Problem of Testing and Estimation Involving Four Fold Tables.**

In this paper, the old problem of the analysis of four fold tables is re-examined with respect to a peculiar kind of sampling. Assuming a population possesses properties A and B, consideration is given to several ways in which samples might be taken in order to determine whether A and B are associated.

Case 1. Simple random sampling. N members of the population are drawn at random and the frequencies of occurrence of A, B; A, not B; not A, B; and not A, not B are observed. This is the classical, simple sampling procedure.

Case 2. Sampling with respect to A. By this is meant the random and independent selection of N_1 members with property A and $N_2 = N - N_1$ members without property A. N_1 and N_2 are fixed in advance and the number of cases of B and not B are observed in each of these two groups.

Case 3 is the reverse of case 2, namely, sampling with respect to B.

In case 4, estimates of various parameters are found by one of Neyman's methods for obtaining best asymptotically normal estimates. A more recent method involving a logistic transformation is also used. The powers of several related χ^2 tests are compared with the result that tests based on sampling by case 2 or 3 above have higher asymptotic power than one based on case 1 sampling. Case 4 lies between case 2 and case 3 in this regard.

715 ROBERT M. THORNER and QUENTIN R. REMEIN (U. S. Public Health Service). **Some Aspects of Screening Tests for the Detection of Disease Suspects.**

The screening of large population groups for chronic disease in a growing health department activity. The purpose of screening is the selection, for diagnostic evaluation, of persons in the population with a high probability of being diseased. The application of screening tests requires a basic knowledge of the attributes of these tests and an understanding of the results that may be anticipated when a particular test is applied to a population group at a selected screening level.

Tests used in screening must be simple and inexpensive; should be sensitive, specific, accurate, and precise. The theory of overlapping distributions is used to explain sensitivity and specificity. The stability of these measures and their relationship to prevalence is discussed. The effect of altered screening levels on sensitivity and specificity is illustrated.

Problems of selecting screening levels are discussed, together with the relationship of sensitivity, specificity and prevalence to the frequency of false negative and positive test results. Youden's test evaluation index is discussed in relation to comparing test results with chance.

The static and stochastic models of screening tests are considered, and "relative" sensitivity and specificity are reviewed. The calculation of confidence limits for sensitivity and specificity percentages is illustrated, and statistical techniques used for comparing results of different screening techniques, for correlated and uncorrelated models, are reviewed and illustrated. "Precision" and "accuracy" are defined, and the measurement of precision and comparison of precision measurements is illustrated.

716 R. M. THRALL (University of Michigan, Willow Run, Mich.). **A Review of Mathematical Aspects of the Theory of Measurement.**

The past fifteen years have witnessed a vast advance in the techniques and theories of measurement. Methods and theories appropriate to the physical sciences have been extensively generalized so as to become more applicable to the behavioral and life sciences. The present paper is limited to a review of certain mathematical systems which either have been used in measurement or show promise of such use.

Measurement in the physical sciences centered on description of some attribute of a physical object by a real number. S. S. Stevens pioneered in generalizing this concept of measurement, and his work stimulated still further generalization and abstraction.

A fundamental point of view in modern measurement theory is that the mathematical system used for measurement should be no stronger than is justified by the properties of the objects being measured. For example, use of the real numbers to measure hardness is not justified since neither addition nor multiplication has meaning relative to hardness. All that is physically meaningful is that relative to hardness objects can be placed in order with ties permitted. Thus for the "degrees of hardness" only a chain order is mathematically justified. The presence of ties in hardness indicates that a weak order is suitable for the objects themselves.

The present paper is devoted to a review of the development during the past fifteen years of mathematical models appropriate for measurement. These models vary in the strengths of the axiom systems used and in the extent to which stochastic theories are employed. Most of the early systems are entirely deterministic; later systems are increasingly stochastic in nature.

717 M. B. WILK and R. GNANADESIKAN (Bell Telephone Laboratories, Murray Hill, N. J.). **Some Remarks on Plotting Procedures in the Analysis of Experiments.**

The paper deals with certain problems of estimation, associated with plotting procedures in analyzing experiments, using rank order statistics. A detailed discussion of the half-normal plotting procedure in this connection is included.

The following are abstracts of papers presented at The Symposium on Quantitative Methods in Pharmacology held in Leyden, The Netherlands, May 10–13, 1960.

718 J. HAJNAL (London School of Economics, London, England). **Sequential Tests on Analgesics in Rheumatoid Arthritis.**

The paper is concerned with a series of clinical trials carried out in the Rheumatism Research Clinic of the Royal Infirmary, Manchester. The background and trial procedure are briefly described. The statistical analysis appropriate to this procedure, if sequential methods are not used, is compared with the sequential method in fact employed. The advantages and disadvantages of incorporating this sequential method in the design are discussed and some possible improvements in the sequential analysis are indicated.

719 N. L. JOHNSON (University College, London, England). **On the Choice of a Sequential Test Procedure.**

1. In the construction of a standard sequential probability ratio test (s.p.r.t.) procedure we need to define (i) two pivotal hypotheses H_0, H_1, (ii) two (approximate) probabilities of error α_0, α_1, such that

$$\Pr \{\text{reject } H_j \mid H_j\} \cong \alpha_j \qquad (j = 1, 2).$$

It is known that the s.p.r.t. is then optimal in the sense that among all procedures satisfying (ii) the s.p.r.t. has (nearly) the smallest expected size of sample whether either H_0 or H_1 is true.

2. Conditions (i) and (ii) are the same as those often used in determining the size of sample to be taken in a classical fixed sample procedure. In such cases the hypotheses H_0 and H_1 (and the probabilities α_0 and α_1) are chosen in the light of the required sensitivity of the procedure. For example H_0 may be a 'null hypothesis' (and α_0 a 'level of significance') corresponding to a desirable level of quality, with H_1 representing a change in quality of such an amount that a high probability $(1—\alpha_1)$ of detecting it is needed.

If H_0 and H_1, determined in this way, are used to define a s.p.r.t., the full advantage of the sequential method may be lost unless either H_0 or H_1 is true (or nearly so) in a large proportion of cases. When a hypothesis H, different from both H_0 and H_1, is true there is no guarantee that the s.p.r.t. procedure defined by H_0, H_1, α_0, α_1 will even have a lower average sample size than a fixed sample procedure defined in the same way.

3. A s.p.r.t. satisfying (ii) need not necessarily be based on H_0 and H_1 as pivotal hypothesis. It is suggested that consideration should be given to the introduction of additional hypotheses H_0' , H_1' , to be used as pivotal hypotheses. H_0' and H_1' would be chosen to represent the sort of conditions expected to be encountered rather frequently—so that the saving in average sample size should be as effective as possible. The associated (approximate) probabilities of error α_0' , α_1' (where Pr {reject $H_j'/H_j'\} \cong \alpha_j'$; $j = 1, 2$) are then determined by conditions (ii).

4. In this paper detailed consideration is given to the case when the s.p.r.t. is regarded as a test of the 'null hypothesis' H_0 , and where we take $H_j' \equiv H_0$. Some tables are given to aid in the construction of the test procedures, and also to describe the properties of such tests.

5. Incidentally to the main topics, some results are obtained on the conparison of operating characteristics (or power functions) of comparable s.p.r.t.'s.

720 CHR. L. RÜMKE (Free University of Amsterdam, Neth.). **An Efficient Design for Comparing the Effects of Two Treatments.**

The variability of the reactions of groups of animals to drugs often interferes with an efficient comparison of the effects of two treatments with all or none responses. In classical ED_{50}-determinations it is difficult to choose such dose-levels that the inefficient 0- or 100% effects are avoided. A design is to be described which leads to an efficient test for the difference between two treatments at an optimal dose level. A rough estimate of the ED_{50}'s can be made by slightly modifying this design—though with some loss of efficiency.

721 K. L. SMITH (Boots Pure Drugs, Nottingham, England). **Sequential Analysis Applied to Biological Control Tests for Pharmacopoeial Substances.**

During some biological examinations to control the quality of pharmacopoeial substances, sufficient evidence may be obtained at one time to satisfy the official requirements and the results from such examinations may be analysed by the traditional methods.

In some examinations, such as those to exclude undue amounts of pyrogens, the evidence collected at one time may be sufficient to accept or reject the material in the extreme cases. For intermediate cases repeat tests are permitted and the criteria for them may be conveniently derived from a sequential sampling plan.

In a large number of tests the evidence must be collected sequentially either as single responses as in the assay of Digitalis using guinea pigs or pigeons, or as the results from repeated tests which individually are too imprecise as in the assay of Insulin. Both circumstances lend themselves to the application of sequential analysis. In both cases where one has set out to make a certain number of observations before applying traditional methods of analysis, the value of using a sequential plan may be one of economy only. In those cases where testing is continued until the sample may be considered satisfactory, it is essential that the amount of repeat testing should be sufficient to satisfy a sequential sampling plan.

722 D. J. MÉWISSEN and E. H. BETZ (l'Universit. de Liège, Belgium). **Sequential Tests in Protective Effects of Cystamine and Chlorpromazine.**

A simple method based on sequential test has been used by KIMBALL and al. (Radiation Research, 7, 1, 1957) for screening compounds in radiation protection.

Under specific conditions, the method is easy to apply and the decision of accepting or rejecting a prospective action of protection is based on the difference in animals surviving in the treated and control groups.

The effectiveness of protective agents widely varies with the delay between administration to animals and exposure. The responsible factors and the underlying mechanisms are poorly understood. Therefore, we investigated variations in the protective action of cystamine and chlorpromazine versus delay in X-irradiation of treated animals. Usual sampling methods would have been unpractical and expensive as existence of a protective effect was questionned under various time specifications. Sequential tests were used. A 60% survival in the treated group was assumed a meaningful increase over 50% survival in the control group. Maximum risks for both kinds of errors were set at 10%. Tests were arbitrarily stopped after five steps. Decision was usually reached after three steps.

In a first series of experiments, groups of ten animals were used. In treatment groups, chlorpromazine (5 mg) was injected intraperitoneally to adult pure inbred rats (average body weight: 180 g), respectively 2, 3 and 4 hours prior to total body irradiation (750 r). In a second series of experiments, groups of 5 animals were used. Cystamine (100 mg/100 g of body weight) was injected intraperitoneally to rats, respectively 2, 10, 20, 30 and 45 minutes prior to irradiation (900 r). In each instance, a sequential test was performed on successive groups. In addition. mortality in all groups was recorded for 30 days and mortality curves were drawn. In all cases, decisions reached at an earlier stage with sequential tests were confirmed by further examination of the completed mortality curves.

723 E. W. PELIKAN (Boston University, Boston, Mass., U. S. A.). **Dose Metameters in Comparative Pharmacology.**

Both inter-species and intra-species comparisons of sensitivity to drugs of common mammals, including man, may be facilitated by expressing dose as a power function of body weight. Determination, from published data, of the constants in the equation [log mean total dose (mg) = b log mean body weight (gm) + a] yielded the power function best suited to inter-specific comparison of potency of drugs of three classes. When paralysis was the criterion of effect, b for tubocurarine, its dimethylether, benzoquinonium, gallamine, and succinylcholine did not differ significantly from 1.0; for decamethonium, b was 0.60, significantly less than 1.0. When anesthesia was the effect, b for amobarbital, pentobarbital and secobarbital was about 0.80 and b was independent of route of parenteral drug administration. When acute death was the effect, b was 0.83 for ouabain, 1.01 for digitalis tincture. In contrast, analogous values of b for digitoxin, digitalis, ouabain or Lanatoside C (computed using data for total lethal doses and body weights of individual members of a given species) varied from 0.52 to 0.89 (median: 0.70) for these drugs in each of several species. Minimizing inter-species variance in potencies of drugs, precision in estimating and predicting drug potencies among species or individuals, and detection of species differences in sensitivity to drugs may be furthered by use of an appropriate dose metameter.

724 P. J. CLARINGBOLD and C. W. EMMENS (University of Sydney, Australia). **Biological Assays Involving Quantal and Semi-quantal Responses.**

Various transforms have been used for interpreting the results of biological assays and other experiments when the response is quantal. Usually, intervals are studied in which transforms such as the probit, logit or angular show no perceptible differences, and it is a matter of practical convenience which is used. The angular transform is often the most useful, as it has equal weights at all response levels.

In much early work, few dose-response lines had to be calculated, and the design of tests was simple. More and more frequently however, screening trials or complex designs require methods which offer rapid analysis, even if approximate. Also, within-animal quantal assay will usually yield few responses from each animal and the application of even the angular transform becomes extremely complex. In such circumstances, analyses of variance of the untransformed data (scored 0, 1 or 0, 1, $\cdots n$) is defensible and gives rapidly computed estimates which show little loss of information and allowed of up to a 4-fold increase in precision in cross-over assays with mice. In factorial tests, probit or even angular transformation leads to formidable computations if many factors are involved, whereas analysis of variance of the untransformed data is simple and remarkably accurate in practice.

725 R. VAN STRIK (Philips-Duphar, Weesp, Neth.). **A Simple Method of Estimating Relative Potency and its Precision from Semi-quantitative Responses.**

A simple procedure is described for evaluating those bio-assay experiments where the observed responses are scored in a semi-quantitative scale, such as =, +, ++, +++. The initial responses are ranked together in order of magnitude and a potency ratio with its confidence limits are calculated from the ranks using customary bio-assay computational methods with just one minor alteration. The method is outlined in an example and its reliability is investigated.

726 CLAUS RERUP (The Royal University, Lund, Sweden). **On the Validity of a Cross-over Assay.**

Cross-over assays are valid only, when the following criteria are fulfilled:

1) The assay design has to be balanced, which means that a set of observations always has to be obtained not only in a certain order, but simultaneously in the reverse order because of possible changes in the magnitude of reactions between treatments within subjects.

2) No significantly different changes in reaction between groups of subjects are allowed to occur from treatment to treatment.

For the twin and triplet-cross-over assay (insulin, corticotrophin) it is shown that the hitherto applied validity tests (in the twin cross-over assay the test for parallelism of the dosage response lines for standard and test, in the triplet cross-over assay extended by a test for curvature of the mentioned lines) are not appropriate, because group differences may cause apparent significant differences in slope. Group differences are, however, eliminated in the cross-over analysis. An example of a twin cross-over assay on insulin is given, which must be regarded as invalid according to hitherto published validity criteria, which, however, is a good and valid assay according to the essential validity criteria to be presented in the paper.

727 J. HEMELRIJK (Technological Institute, Delft, Neth.). **Experimental Comparison of Student's and Wilcoxon's Two Sample Tests.**

The choice between the many available statistical methods is often difficult to make in practical situations. It is often—too often—based on personal preference without sufficient regard for the merits of the different methods available. A lack of knowledge about the small sample properties of many distribution-free methods tends to make their position uncertain in comparison to classical methods, based on specified suppositions about the sampled populations; the properties of the latter are known, if these suppositions are true.

In a small experiment, using QUENOUILLE's (Biometrika, 46 (1959), 178—204) 50 independent pairs of samples of size 10 were used to compare STUDENT's and WILCOXON's two sample tests.

The samples were first taken from a standard normal distribution (QUENOUILLE's $_1$), then one of them was shifted over distances of 0.81; 1.05 and 1.53 to the right. These shifts correspond to a power for STUDENT's test of 0.40; 0.60 and 0.90 respectively. In all these positions both tests were applied onesidedly with level of significance 0.025. As was to be expected—STUDENT's test being uniformly most powerful in this situation—the power of WILCOXON's test was the smaller one. The results indicated that in this case one would not be far off the mark in stating that, roughly, the loss of power was about 10% of the power of STUDENT's test as long as the latter is not too close to 1. (This also holds under the hypothesis tested: the true level of significance is for STUDENT's test 0.025 and for WILCOXON's test, owing to its discrete character, 0.022 in this case.)

A second point of investigation was the reaction of the tests to an observational error in one of the original observations. For each pair of samples an observation was chosen at random from the sample with the larger median and this observation was shifted over a distance 2 to the right. In accordance with expectation STUDENT's test proved to be far more sensitive to this mild kind of slippage than WILCOXON's test.

Finally all observations, transformed to observations from a rather skew exponential distribution (QUENOUILLE's χ_4), were again tested in the same way. The shifts were in this case replaced by multiplication of the observations of one of the samples from the starting point of the distribution, in order to keep this point invariant. The factors were: 1 (identical distributions); 2; 3 and 4.5, resulting in power function values of roughly 0.025 (the level of significance); 0.25; 0.45 and 0.70 respectively. The two tests were now approximately equally powerful.

All these results are dependent in a statistical sense, being based on the same 1000 observations. Also the investigation was too small (although laborious) to warrant strong conclusions. It is only the beginning of more extensive research with more powerful means.

H. DE JONGE (Netherlands Institute of Preventive Medicine, Leyden, Neth.).

728 **The Influence of Non-normality on the Significance Levels of Student's Two Sample Test.**

The paper describes an empirical sampling study of the distribution of Student's t for small samples taken from four populations with non-normal distributions, respectively showing: A. Marked skewness, B. Skewness with a bunch of outliers, C. Leptokurtosis with skewness, D. Platykurtosis with skewness. These types of distributions were chosen, because they were frequently seen in practice.

From each population pairs of random samples of 5, 10 and 20 elements were drawn and the values of t were computed. For each of the populations and for each sample size the empirical sampling distribution of t has been compared with the corresponding Student-distribution.

729 H. R. VAN DER VAART (University of Leyden, Neth.). **On the Robustness of Wilcoxon's Two Sample Test.**

It has long been known that the probability of Type I error of Student's two sample test is subject to appreciable variation in the case of unequal population variances (Hsu, 1938, Stat. Res. Memoirs; Chand, 1950, Ann. Math. Stat.). Therefore it is interesting to know the behaviour of Wilcoxon's two sample test in this respect. Van der Vaart (1950), Indag. Math.) found in a special case for very small samples that Wilcoxon's test is less sensitive to inequality of population variances than is Student's tests, Kruskal and Wallis (1952, Journ. Amer. Stat. Assoc.) put forward an argument which lends plausibility to the conjecture that Wilcoxon's test may be fairly insensitive to differences in variability, and Hodges and Lehmann (1956, Ann. Math. Stat. p. 327–329) show that Wilcoxon's test is often less powerful against contamination with a shift than is Student's test. All this might lead one to expect that Wilcoxon's test would in general be less sensitive to the above-mentioned inequality of variances than is Student's test. Now this turns out not to be uniformly true. For instance, in the extreme case where one of the population variances is zero, the relation between the sample sizes determines which test will behave worst in this respect. In the general case the situation is rather complicated.

730 I. JUVANCZ (Math. Institute, Hungarian Academy of Science, Budapest, Hungary). **Contra Indication of Non-parametric Tests in Medical Experimentation.**

According to the author's personal experience from the design and/or analysis of over 700 experiments (most of them medical), the statisticians engaged in medical experiments show a reluctance to use non-parametric methods, though numerous theorists enthusiastically advocate them. A statistical survey is made of papers from some medical, biometrical and statistical periodicals, published in various countries. The results show the same tendency as the author has noted at home.

The reluctance of those "qualified in medicine or, if not so qualified, at least well soaked in it" is caused by some characteristic features of medical experimentation, well known to them. So e.g. in medical experiments the interest is mainly centered on the estimation of mean values and their fiducial limits, and on material significance. There is great variation between and within individuals; the measurements are poor; the number of observations is small. Experiments are sometimes dangerous. Numerous factors disregarded in the statistical analysis influence the final decision. Experimental results are retested several times by others before they are accepted. The results are often intended for individual use (e.g. diagnostic test). Etc. etc.

These characteristics compel the researchers to use statistical methods which yield results convertible to the original measure (e.g. kg, cm). Order statistics and the χ^2-tests are not of this kind. Further, to minimize the loss of information non-parametric methods are too wasteful. Consequently non-parametric tests are economically inefficient both qualitatively and quantitatively.

Not rarely are non-parametric tests insensitive to changes regarded medically substantial and too sensitive to insubstantial ones. (E.g. χ^2-tests are insensitive to consistent though little increasements and sensitive to great though inconsistent ones; this is not the case with the t-test.) If is often impossible to decide what caused the statistically significant change indicated by a non-parametric test (mean, type of distribution etc.). On the other hand, tests of the mean are very "robust" against non-normality.

Only s small field of application is left for non-parametric methods: the cases where the experimental techniques are too weak; e.g. when the results are only given in rank-numbers as sometimes in psychometrics.

Consequently, the use of non-parametric methods should be avoided, if necessary by the invention of better experimental (e.g. chemical) methods.

731 L. J. GOLDBERG (University of California, Berkeley, Cal., U. S. A.). **A Visual Statistical Approach to Bio-assay.**

In quantal bio-assay one usually starts with the following information:

'A series of doses ($d_1, d_2, \cdots, d_i$) has been administered to ($N_1, N_2, \cdots, N_i$) test hosts respectively, and the corresponding reactors ($r_1, r_2, \cdots, r_i$) are observed at each test dose level'.

Based on a set of such experimental observations, a visual procedure will be presented for estimating the true dose response. The proposed procedure is based on the concept that a single experimental observation should be considered as a triplet (r, N, d). The entire set of experimental observations may be considered as follows, in triplet notation.

$$(r_1, N_1, d_1), (r_2, N_2, d_2), \cdots, (r_i, N_i, d_i).$$

The problem of estimation is reduced to the operation of 'triplet addition' conditional to an assumed dose response model. The rules for triplet addition will be illustrated in detail for the single hit dose response model. The extension to other dose response models will be considered, including the logit and probit as specific examples.

732 CONSTANCE VAN EEDEN (Mathematical Centre, Amsterdam, Neth.). **On Distributionfree Bio-assay.**

In bio-assay the following situation is considered: a stimulus (e.g. a drug or a vitamin) is applied to a subject, resulting in a response produced by the subject. In this paper we only consider the quantal (all-or-nothing) response. Then for each subject there will be a level of intensity of the stimulus above which the response occurs and below which it does not occur; this level is called the tolerance of the subject. On a population of subjects this tolerance will be a random variable λ, with distribution function $F(\lambda)$, say. The problem in bio-assay is to give an estimate of this distribution function $F(\lambda)$. A second problem is to compare the distribution functions $F_1(\lambda)$ and $F_2(\lambda)$ for two different stimuli. The observations then consist of the results of applying the stimulus once to each of a number of subjects, at several doses.

The abovementioned problems are solved for the parametric case, e.g. when λ has a normal distribution; the methods given for this situation are called "probit analysis" and are described e.g. by D. J. FINNEY (1947).

The application of this method, however, requires in some cases laborious computational work. Moreover difficulties arise if, for one or more doses the number of observations is very small or if none of the subjects or all subjects give a response. Finally the assumption of normality may not be fullfilled.

This paper contains a description of what may be obtained if no assumptions are made on the form of the distribution of λ. Let observations be available at

k different doses $d_1, \cdots, d_k$ with $d_1 < \cdots < d_k$. The number of observations at dose d_i is denoted by n_i and the number of subjects giving a response at dose d_i is denoted by a_i. We suppose all observations to be independent. Let further p_i denote the probability that a subject gives a response at dose d_i then $p_i = F(d_i)$; consequently, $F(\lambda)$ being a monotone non-decreasing function of λ, we have $p_1 \leqq \cdots \leqq p_k$.

For this situation maximum likelihood estimates of $p_i = F(d_i)$ may be obtained by means of a method described in the thesis of C. VAN EEDEN (1958). In this thesis a more general problem is considered, where k parameters of k distribution-functions are to be estimated, if it is known that these parameters are partially or completely ordered and moreover confined to given intervals I_i. If the distributions are binomial, the ordering is complete and I_i is the interval [0, 1], we obtain the above mentioned problem of estimating probabilities $p_1, \cdots, p_k$ satisfying $p_1 \leqq \cdots \leqq p_k$.

These maximum likelihood estimates may be obtained in a simple way and it is not necessary to have a large number of observations at each dose. Thus a maximum likelihood estimate of $F(\lambda)$, at the doses used in the experiment, is obtained without any assumptions of a parametric nature about the distribution function $F(\lambda)$.

References

VAN EEDEN, C. (1958), Testing and estimating ordered parameters of probability distributions, Thesis Amsterdam.

FINNEY, D. J. (1947), Probit Analysis, Cambridge University Press, Cambridge.

733 C. W. DUNNETT (University of Aberdeen, Scotland). **Statistical Theory of Drug Screening.**

Screening a chemical compound for some specific form of chemotherapeutic activity is considered as a decision problem in which the terminal decisions open to the investigator are either to accept the compound as being worthy of further investigation or to reject it as being of no interest and the unknown "state of nature" is the degree of activity of the compound. A truncated sequential procedure for screening is proposed, in which rejection can occur at any stage but acceptance is allowed only at the final stage. The particular case in which only two levels of activity exist (called "active" and "inactive") and all compounds have the same *a priori* probability of being active is considered in detail. A criterion of optimality is introduced, which involves the cost of testing and the costs associated with wrong decisions. A method for computing the critical rejection levels when the testing errors are normally distributed is given, and illustrated for one, two and three-stage procedures. Extensions to cases in which some compounds are more likely to be active than others, and in which more than two levels of activity may occur, are discussed.

734 M. A. SCHNEIDERMAN (Cancer Research Institute, Bethesda, Md., U.S.A.). **Statistical Problems in the Search for Anti-cancer Drugs by the National Cancer Institute of the United States.**

The development of the external (contract) screening program of the Cancer Chemotherapy National Service Center is described. The basic experiment is a joint toxicity-therapeutic trial. This permits the administration of maximum tolerated doses without the need for extensive prior toxicity testing. The basic screen uses three different mouse tumors with each tumor system examined independently. A multi-stage sequential scheme is exployed; a two-stage scheme for "natural" products and a three stage scheme for synthetic chemicals. The operating characteristic curves for the two systems are compared. The acceptance and rejection levels have been so adjusted that the two systems yield the same number of false positives, while the two-stage scheme yields more false negatives. This can be tolerated because of the nature of the natural products coming to the screen. The actual behavior of one of the screens is appraised by examining the responses to a "known positive" compound, tested by different screeners in different places at different times.

Problems of mechanical data processing, the use of additional screens beyond the basic three, secondary screening, optimal dose-finding, "positive" and "negative" controls, possible deterioration of natural products, and some possible nonparametric approaches are considered, as examples of unsolved problems.

735 J. J. GRIMSHAW and P. F. D'ARCY (Allen Hanburys, Ltd., Ware, England). **Some Problems Arising in Drug Standardization and the Selective Screening of Compounds of Potential Pharmacological Interest.**

During the course of routine standardization and screening procedures a variety of problems have been encountered. Examples of *in vivo* and *in vitro* techniques are given, in which the assays of purgatives and succinylcholine together with the screening of potential analeptics and barbiturate potentiating drugs, antihistamines and neuromuscular blocking agents are considered. Solutions to some of the problems are proposed and the need for careful statistical control throughout is emphasised. It is suggested that, in some cases, insistence on statistical rigour may lead to results of diminishing practical importance, and the implications of this are discussed.

736 NORBERT BROCK (Asta Werke, Brackwede, Germany) and BERTHOLD SCHNEIDER (Karlsruche, Germany). **Pharmacological Characterization of Drugs by Means of the Therapeutic Index.**

For the pharmacological characterization of drugs it is not sufficient to describe their different qualities of action as such; in order to obtain a clear picture of the value and dangers of any substance, i.e. of its margin of safety, it is necessary to plot one of its actions against the other. As a measure of the margin of safety we introduced the therapeutic index (*DL* 5/*DC* 95 or *DL* 5/*DE* 95) which expresses the relation between the curative and toxic doses. Based on more than 10 years' experience, the author discusses some practical examples in various fields of pharmacotherapy. The trouble in applying the therapeutic index is the difficulty in obtaining approximate values for the deviations of the index and in having suitable testing methods available. Both problems are discussed and suggestions made for practical use.

737 E. J. ARIËNS (University of Nijmegen, Neth.). **Analysis of the Action of Drugs and Drug Combinations.**

As a rule the determination of biological activities of drugs aims at the comparison of the results obtained with various drugs or with one drug under different circumstances, as for instance different doses. The experiments must therefore be based on a sound design and the experimental results evaluated with the aid of statistical methods. A primary question remains, however, whether the effects or magnitudes that are compared are really comparable in the sense meant by the investigator. This is especially of importance if the experimental results serve the study of relations between the chemical structure of drugs and their biological activity. Based on the result of such studies a more direct design of new drugs, with a minimum of trial and error in the research procedure, might be possible. If experimental data are compared, an analysis of the biological effects studied is necessary in order to find out if, to what degree and in what sense these data are comparable. This will be demonstrated by some examples.

738 S. DIKSTEIN (Hadassah Medical School, Jerusalem, Israel). **Physical Chemistry of Drug-receptor Interaction.**

1) The effectivity in vitro of homologous series of aliphatic choline esters, straight chain aliphatic primary amines and straight chain aliphatic monoquaternary ammonium compounds was determined on the small intestine of guinea pig.

2) Morphine, an inhibitor of the acetylcholine release, inhibits the contractions caused by small doses of nicotine, but has no effect on the contraction caused by acetylcholine. This phenomena is similar to the inhibition caused by botulinum toxin, which is known to act at the nerve ending. Since morphine has no effect on the contractions caused by any of the homologous amines tested by us, it was concluded that the primary amines and monoquaternary ammonium ions act directly on the smooth muscle.

3) In the homologous series, the activity increases with increasing number of methylene groups up to a certain limit, the further addition of a single methylene group then abolishes the contractive power or makes the compound inhibitory. Addition of more methylene groups first increases the inhibitory power and then decreases it.

4) A qualitative explanation is offered to explain the observed facts. According to this hypothesis in the presence of a positive charge, if the solubility parameter of the compound is higher than that of the biophase we get excitation, if the solubility parameter is lower we obtain inhibition. The smaller the difference, the greater the effectivity. On the molecular level the excitation is explained by other workers by the interaction of materials with monolayers. According to this work it is postulated that if the material has a lower solubility parameter than that of the biophase it will dissolve in it increasing the spreading force. A material with higher solubility parameter will decrease the spreading force.

5) Since the excitation is a kinetic phenomenon (the velocity of addition of the excitant determines the activity) whereas the solubility parameter represent in fact a term of energy, it is not difficult to explain the absence of quantitative predictions.

6) The importance of the solubility parameter in other drugs is discussed.

739 S. E. DE JONG (University of Leyden, Neth.). **Isoboles.**

In this lecture reference is made to Loewe's method, by which the pharmacological effect of two drugs in combination is diagrammatically represented by means of isoboles.

A definition with a short description is given, followed by a survey of the most important types. Some special cases are discussed more fully. It is explained that some forms of synergism do not lend themselves to being represented by means of isoboles (allobiotic synergism; inversion). Not only because of this do the isoboles not yield an optimal solution to the pertaining problem. It may be considered objectionable that it is necessary to estimate the exact amounts of a series of mixtures of two drugs needed for a quantitatively specified effect. Besides, the isobole thus obtained even then gives only incomplete information; for stronger or weaker effects an entirely different type of isobole can be found.

In conclusion Loewe's rather complicated nomenclature is enumerated.

740 P. S. HEWLETT and R. L. PLACKETT (Pest Infestation Laboratory, Slough, England). **Models for Quantal Responses to Mixtures of Two Drugs.**

The various previous approaches to the construction of mathematical models for quantal responses to mixtures of drugs are briefly referred to, and it is pointed out that models for graded responses to mixtures cannot in general be transferred direct to the interpretation of quantal responses. A general method for developing models for quantal responses to mixtures is put forward. This is exemplified by development of a model for simple similar joint action, i.e. a situation in which two drugs have a common site of action, but in which neither modifies the behaviour of the other at their site of action or elsewhere.

741 J. GURLAND (Iowa State University, Ames, Iowa, U. S. A.). **Determination of Minute Insecticidal Residues through Biological Assay.**

Samples of alfalfa from a field sprayed uniformly with Guthion, were collected 1 day, 3 days, 7 days, 2 weeks, 4 weeks after spraying. The residue in these samples was determined by means of a biological assay in which houseflies were exposed to increasing doses of the insecticide. Since sample extracts contained lipids and other materials which caused a masking effect, the standard preparations used for comparison were made to contain comparable amounts of control extract. The usual type of parallel line assays employing mortality of houseflies were employed for determining the residues in the 1-day, 3-day, and 7-day samples, but the extracts for the later samples had to be fortified with known amounts of the Standard in order to raise the response from a very low level to a level at which a parallel line assay could again be employed.

To increase the sensitivity of the technique it was found that the ratio of the volume containing the fortifying Standard to the volume of the Test extract should be made as small as possible. Further, the final volume of Test extract should be small relative to the sample mass being extracted in order to increase the sensitivity expressed in such units as parts per million of sprayed alfalfa.

742 A. G. MATHEWS (Dept. of Health, Victoria, Australia). **The Interpretation of Antibiotic Blood Level Curves.**

Although hundreds of investigations of antibiotic blood levels have been reported, in comparatively few has any examination of the statistical significance of observed differences been performed, and in even fewer has any attempt been made to assess the pharmacological significance of the results. Consequently, the literature is filled with conflicting reports as to the relative efficacy of various dosage forms, and the lack of adherence to any common experimental design makes the resolution of such conflicts very difficult.

The problem of deriving a single statistic, with which to summarize the information contained in a set of blood level curves, is rendered difficult through ignorance of the relative importance of the height of the peak, the time at which the peak occurs, and the period during which the curve is above an arbitrarily chosen level. Attempts which have been made to derive such statistics will be reviewed, and their limitations discussed.

The shape of the blood level curve in the region of the peak is usually difficult to determine, because of practical objections to the frequent collection of samples of blood. Advantages of experimental designs, in which only a few samples are required from each of many subjects, will be discussed.

No matter how carefully sets of blood level curves may be determined, nor how ingeniously their properties may be summarized statistically, the problem of pharmacological interpretation remains. Stress will be laid on the importance of ancillary studies: for example, of the rate of excretion of the antibiotic, and of its distribution in the tissues of the body.

743 W. Z. BILLEWICZ (University of Aberdeen, Scotland). **A Statistical Investigation of Factors Affecting the Accuracy of D_2O Determinations by the Falling Drop Method.**

After a short description of the method the results of the analysis of the trial series are briefly discussed and taken as the basis of subsequent experiments. The experimental results cover various aspects of the method. The effects of changes in room temperature, position of the dropping tube and dropping technique are considered. Errors of the distillation stage are considered and discounted. The problem of the optimal selection of tube and drop size is considered in some detail in terms of an index of sensitivity of the method. Errors resulting from observer timing as opposed to photo-electric timing are discussed. A comparison of two main estimation methods is made and finally the problem of serial estimates on the same subject is considered.

744 J. DUFRENOY (Conservatoire National des Arts et Métiers, Paris, France). **Using the Arc Tangent Scale for Quantitative Methods.**

The arc tangent scale, used either for the abscissa or the ordinate scale, or for both, is most convenient for plotting "responses" as the dependant variable, in the range 0 to 100, against the independant variable, in the range 0 to infinity.

The arc tangent is most eminently suited for transformation of the time (t) into a metameter (τ).

THE BIOMETRIC SOCIETY

British Region

The following Officers have been elected for 1961:

President: J. A. Fraser Roberts
Secretary: C. D. Kemp
Treasurer: P. A. Young.

The Summer Meeting of the Region was held on July 6, 1960, at the National Vegetable Research Station, Wellesbourne, Warwick.

A meeting on October 27, 1960, was concerned with "Applications of Electronic Computers to Biological Problems." The following papers were read and discussed:

W. T. Williams—Some Applications of Electronic Computers in Plant Ecology and Taxonomy.
J. N. R. Jeffers—Data Processing in Biological Research.
F. Yates—Problems Arising in the Use of Electronic Computers in Statistical Analysis.

Following the Annual General Meeting on December 13, 1960, the following papers were read and discussed:

C. W. Emmens—The Planning and Analysis of Some Field Trials with Cattle.
C. C. Spicer—Problems in the Analysis of a Large Scale Clinicial Trial.

On January 2, 1961, a special meeting was held on "Biometrical Aspects of Plant Growth" in collaboration with the Society for Experimental Biology. The programme included:

J. A. Nelder—Models and Experiments for Growth Analysis
S. C. Pearce and C. S Moore—A Study of the Sources of Variation in Growth of Fruit Trees.
M. J. R. Healy—Experiments for Comparing Growth Curves.

A general discussion was led by F. L. Milthorpe.

E.N.A.R.

The following officers have been elected for 1961 by the Eastern North American Region:

President, 1961: Oscar Kempthorne
President-elect, 1961: Henry L. Lucas
Secretary, 1961: Erwin L. LeClerg
Treasurer, 1961: Donald A. Gardiner
Regional Committee, 1961–63: Virgil L. Anderson and Robert J. Monroe.

ENAR TREASURER'S REPORT FOR 1960

Marvin A. Kastenbaum, *Secretary-Treasurer*

INCOME

Balance Forward	
Checkbook balance Dec. 31, 1959	$ 324.29
Dues payments	5,181.50
Share of Proceeds from Washington meeting (1959)	267.88
Credit on Canadian checks	.40
Payment from WNAR for Programs	5.00
Sustaining member credit	100.00
Payment from AIBS for printing	14.00
	$5,893.07

EXPENSES

Checks not honored		$ 14.00
Refund on dues		.50
Treasurer International		4,465.00
Transfer to Savings Account		267.88
ENAR expenses		
Postage	$187.93	
Printing	365.78	
Clerical	30.96	584.67
Checkbook Balance December 31, 1960		561.02
		$5,893.07

OPERATING ANALYSIS

Operating income		
Dues payments	$702.00	
Other societies	19.00	
Bank credit on checks	.40	
Sustaining member credit	100.00	$ 821.40
Operating expense		584.67
Surplus for 1960		$ 236.73

CHECKBOOK BALANCE

Advanced from 1959	324.29	
Surplus for 1960	236.73	
Checkbook balance November 15, 1960		$ 561.02

SAVINGS ACCOUNT BALANCE

Advanced from 1959	716.11	
Interest 1959	6.26	
Proceeds from Washington meeting (1959)	267.88	990.25
Total cash in Bank December 31, 1960		$1,551.27

W.N.A.R.

MINUTES OF THE WNAR BUSINESS MEETING

The annual meeting of the Western North American Region of the Biometric Society was held at 1:00 P.M., August 24, 1960, on the Stanford University Campus. William Taylor, President, WNAR, presided. The highlights of the minutes of the previous meeting were read by the Secretary, Walter Becker, and were approved. The Secretary then gave his report. A method for obtaining News and Announcement for *Biometrics* was described. This consisted of sending to everyone who changed his address an inquiry as to his position, etc. This system, suggested by the ENAR Secretary has produced excellent results. Sometimes ballots are sent to student members inadvertently and a reminder was made that they are not entitled to vote. Four general mailings to the members of WNAR have been made this year The postage cost is about $6 a mailing, but this figure may go up as a result of the Secretary's office being moved to Washington State University, Pullman, Washington.

The Treasurer, Bernice Brown, gave her report. WNAR now has 164 members, 8 of them students, an overall gain of 14 from last year. As of August 20, 1960, there was $445.25 on hand. The operating expenses balanced out the income for this year. It was brought out that much of the money in the treasury was accumulated during past administrations when much of the cost of printing and mailing was not born by the Society. WNAR receives $1 for every member, $10 for each sustaining member, and nothing from student members. The Treasurer stated that perhaps a working capital of $200 would be sufficient in the treasury, declaring $200 to be surplus.

The Nominating Committee made its report. Two positions on the Regional Committee are now open because of the expiration of the terms of Douglas Chapman and Donald Wohlschlag. The names submitted by the Nominating Committee were Donald Owen, New Mexico; Lincoln Moses, California; Marion Sandomire, California; and Douglas Chapman, Washington. No nominations were made from the floor. The vote will be taken by a mail ballot.

No action was reported by the Relationship of ASA and the Biometric Society Committee during the past year by Marion Sandomire. The By-laws of the region are being revised by William Taylor and will be submitted soon to the Regional Committee. The possibility of having a president-elect was discussed. At present the President is elected for two years, in order that he may be more experienced in the position. Several alternatives were proposed:

1. The presidency be made a one-year position. Then we would have a president-elect every year. The president-elect would act as a program chairman and would in addition receive copies of important letters from the officers and the Regional Committee.
2. On alternate years, a president-elect would be elected.
3. Leave as is.

It was pointed out that the need for a president-elect might become greater as the membership increased. It was suggested that this matter be considered by the Regional Committee and be submitted to the membership in a mail ballot.

Douglas Chapman moved and Calvin Zippin seconded the following motion: The By-laws may be amended by a two-thirds vote of members present at any annual meeting. Changes in the By-laws may be initiated by the Regional Committee and the By-laws amended by a majority of members voting in a mail ballot. This motion was passed by a unanimous vote.

Suggestions for the use of the surplus fund (about $200) in the treasury were:

1. Give prizes for the best student paper presented at an annual meeting (it was mentioned that this was tried a couple of years ago, but that students did not present any papers).
2. Do not use the money.
3. Send bound volumes of *Biometrics* to foreign universities.
4. Pay students' dues.
5. Pay the expenses of a speaker at the annual meeting.
6. The funds should be used to encourage interest in the society.
7. They could be used to promote membership (it was pointed out that new members' dues would increase the treasury's surplus).
8. Prizes could be given to the member bringing in the most new members.
9. A year's subscription could be given to outstanding students.

No agreement could be reached, so the Regional Committee will debate the issue by mail.

Two possible meeting places for the 1961 Annual Meeting were discussed. These were the meeting of the Pacific Division A.A.A.S. at the University of California, Davis, June 12–17, 1961 and the IMS Meeting at the University of Washington, Seattle, in August (later information indicates that this may be in June). A majority voted for the Seattle meeting.

The purpose of the program at the Annual Meeting was discussed. Should the program aim at a specific purpose, or should the papers be an expository review aimed at applied statistics? There was very little comment.

William Taylor announced that he would be attending a meeting in the East with representatives of other societies to consider the possibility of a federation of statistical societies, besides other purposes.

The meeting adjourned at 2:10 P.M.

Those present at the meeting were Abramson, Becker, Bennett, Brown, Chapman, Hopkins, Massey, Mode, Nash, Nicholson, Russell, Sandomire, Taylor, Vaughan, and Zippen.

NOTE ON ELECTION

Douglas G. Chapman and Marion M. Sandomire were elected to the Regional Committee for 1961–1963.

CHANGES IN MEMBERSHIP

(October 1, 1960–January 15, 1961)

Changes of Address

Miss Margaret F. Allen, Department of Biometrics, School of Aviation Medicine, USAF, Brooks Air Force Base, Texas, U. S. A.

Dr. David W. Alling, Institute Allergy and Infectious Disease, National Institutes of Health, Bethesda, Maryland, U. S. A.

Professor Maurice S. Bartlett, Department of Statistics, University College, London W. C. 1, England.

Prof. Dr. Eduard Batschelet, 2500 Q Street N.W., Washington, D. C., U. S. A.

Mr. Rainald K. Bauer, Kolner Str. 228a, Dusseldorf, Germany.

Dr. Gilbert W. Beebe, National Academy of Sciences, 2101 Constitution Avenue, N.W., Washington 25, D. C., U. S. A.

Dr. G. E. P. Box, Mathematics Research Center, U. S. Army, University of Wisconsin, Madison 6, Wisconsin, U. S. A.

Dr. David Bruce, Box 4059, Portland, Oregon, U. S. A.

Dr. Joseph G. Bryan, 11 Newport Avenue, W. Hartford, Connecticut, U. S. A.

Mr. Walter E. Cole, Division of Forest Insect Research, U. S. Forest Service, Ogden, Utah, U. S. A.

Mr. Jerry Cornfield, National Institutes of Health, Bethesda, Maryland, U. S. A.

Dr. Paul M. Denson, Deputy Commissioner of Health, 125 Worth Street, New York 13, N. Y., U. S. A.

Prof. Benjamin Epstein, 768 Garland Drive, Palo Alto, California, U. S. A.

Dr. Robert Flamant, 11, rue Bouille, Fontenay—Aux—Roses (Seine) France.

Dr. William A. Glenn, 906 King Street, Cary, North Carolina, U. S. A.

Prof. John Gurland, U. S. Army Research Center, University of Wisconsin, Madison, Wisconsin, U. S. A.

Dr. Thomas J. Haley, 3401 Woodcliff Road, Sherman Oaks, California, U. S. A.

Dr. Sudako Hayase, 1038 Fifth Street, Santa Monica, California, U. S. A.

Prof. Dr. Ottokar Heinisch, Grunderstr. 28, Berlin—Grunau, Germany.

Dr. Henry Hopp, Agricultural Attache—Bogota, c/o U. S. Department of State, Washington 25, D. C., U. S. A.

Professor Gwilym Jenkins, Department of Mathematics, Imperial College, London, S.W. 7, England.

Dr. S. Karatas, Department of Animal Husbandry, University of Ataturk, Erzurum, Turkey.

Dr. Eileen B. Karsh, Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania.

Mr. Charles Kiss, La Menitre, (Maine-et-Loire) France.

Mr. Gary F. Krause, 1110 Kentwood Drive, Blacksburg, Virginia, U. S. A.

Mr. William T. Lewish, 2109 Wildwood Drive, Woodland Park, Wilmington 5, Delaware, U. S. A.

Mr. Donald G. MacEachern 608 Third Avenue, S. E., Minneapolis 14, Minnesota, U. S. A.

Mr. John W. Mayne, Van Zaeckstraat 49, The Hague, Netherlands.

Mr. Jean Mothes, 13, boulevard des Invalides, Paris 7e, France.

Mr. Clifford A. Myers, 2069 N. Navajo, Flagstaff, Arizona, U. S. A.

Dr. Wesley L. Nicholson, 1215 E. 3rd, Moscow, Idaho, U. S. A.

Dr. Fernando Orozco-Pinan, Explotaction "El Encin" Alcala de Henares (Madrid) Spain.

Dr. Laurence M. Potter, Department of Poultry Husbandry, Virginia Polytechnic Institute, Blacksburg, Virginia, U. S. A.

Dr. D. R. Read, Cadbury Brothers Limited, Bournville, Birmingham, England.

Mr. A. H. L. Rotti, Av. Emile Verhaeren 70, Gentbrugge, Belgium.

Mlle. Claude Rouquette, 3, rue Saint-Charles, Paris 15, France.

M. Jean Soule, 30 avenue Saint-Laurent c. 34 (Orsay) Seine-et-Oise, France.

Mr. John J. Sowinski, Allstate Insurance Company, 7447 Skokie, Blvd., Skokie, Illinois, U. S. A.

Prof. David F. Votaw, Jr., 6 Fairlane Terrace, Winchester, Massachusetts, U. S. A.

Mr. Jerry Warren, Department of Horticulture, University of Nebraska, Lincoln 3, Nebraska, U. S. A.

Mr. Francis R. Watson, 1415 Winding Lane, Champaign, Illinois, U. S. A.

Mrs. Sandra S. White, 535 E. 72nd Street, New York 21, N. Y., U. S. A.

Mr. Amador D. Yniguez, Institute of Statistics, P. O. Box 5457, Raleigh, North Carolina, U. S. A.

Dr. Sydney S. Y. Young, 68 Barker Road, Strathfield, N.S.W., Australia.

New Members

At Large

Mr. Pedro Eliz M. Revello, Instituto Venezolano Petroquim, Centro Simon Bolivar, Caracas, Venezuela.

Brazil

Mr. Joassy de Paula Neves Jorge, Instituto Agronomico Caixa Postal 28, Campinas S.P., Brazil.

France

M. Pierre Lossois, 120 bis avenue de Verdun, Issy-les-Moulineaux (Seine) France.

Germany

Dipl. G. Enderlein, Karl-Marx-Str. 31, Klein Wenzleben bei, Magdeburg/DDR, Germany.

Mr. Werner Haufe, Sophienstr. 21, Einbeck/Hann., Germany.

Prof. Dr. A. Maede, Eichenweg 5, Halle/Saale, Germany.

Mr. Dieter Rasch, Fritz Reiter Str. 22, Rostock, Germany.

Dr. Ernst Weber, Windhalmweg 36, Stuttgart-Plieninger, Germany.

Doz. Dr. Hermann Witting, Dreikonigstrasse 9, Freiburg i. Br., Germany.

Italy

Dott. Giuseppe Agnese, Istituto di Igiene dell'Universita, via Pastore 1, Genova, Italy.

Dr. Giancarlo Chisci, Stazione Sperimentale di Praticoltura, Viale Piacenza 25, Lodi, Milano, Italy.

Dr. Maurizio Turri, Viale O. P. Vigliano 17, Milano, Italy.

Western North American Region

Mr. Peter A. Dawson, Department of Genetics, University of California, Berkeley 4, California, U. S. A.

Dr. Olive Jean Dunn, School of Public Health, University of California, Los Angeles 24, California, U. S. A.

Prof. Robert F. Fagot, Department of Psychology, University of Oregon, Eugene, Oregon, U. S. A.

Dr. Robert Macey, Department of Physiology, University of California, Berkeley 4, California, U. S. A.

Mr. Prem Singh Puri, 1845 Hearst Avenue, Berkeley 3, California, U. S. A.

Mr. Kurt Sittman, Animal Husbandry Department, University of California, Davis, California, U. S. A.

Mr. Albert R. Stage, 157 South Howard Street, Spokane 4, Washington, U. S. A.

NEWS AND ANNOUNCEMENTS

Members are invited to transmit to their National or Regional Secretary (if members at large, to the General Secretary) news of appointments, distinctions, or retirements, and announcements of professional interest.

EDITORIAL NOTE

The general objective of *Biometrics* is to promote and to extend the use of mathematical and statistical methods in pure and applied biological sciences. This objective has been broadly interpreted in editorial decisions and this is proper. However, it is believed also that mathematicians and statisticians contributing to *Biometrics* have an obligation to present their work so that the natures of the problems in the biological sciences to which the methods apply are clearly understood by those who may profit by the use of the methods. Similarly, biologists should explain their problems so that their needs and methods are clearly understood by the mathematician and statistician. While these requirements may not be easy to meet, direct consideration of them should greatly increase the value of *Biometrics* to all members of the Biometric Society. Careful attention to introductions of papers, including formulation of problems and models, with skillful selection of examples would be helpful. When possible, algebraic and mathematical discussions should be placed in appendices to papers in order that they do not detract from the emphasis on the problem areas and the readability of papers.

MEETINGS OF E. N. A. R.

The Eastern North American Region will meet jointly with the Institute of Mathematical Statistics and the American Statistical Association on April 20, 21, and 22, 1961, at Cornell University. Titles and abstracts, the latter in duplicate in the form published in *Biometrics*, of contributed papers for ENAR should be sent to Dr. Erwin L. LeClerg, Biometrical Services, Plant Industry Station, Beltsville, Maryland.

In 1961, ENAR will also meet jointly with the American Institute for Biological Sciences at Purdue University and with the American Statistical Association in New York City.

MEETING OF W.N.A.R.

The Western North American Region will meet jointly with the Institute of Mathematical Statistics in June, 1961, at the University of Washington.

UNIVERSITY ANNOUNCEMENTS

THE CATHOLIC UNIVERSITY OF AMERICA

The Statistical Laboratory of the Catholic University of America has been awarded a grant by the National Institutes of Health for training in the field of Biometry.

The stipends for first-year graduate students are $2250 plus tuition; family allowances for dependents and annual increases are provided.

The students will pursue the same general program as other students in mathematical Statistics. They will participate in the consulting activities of the laboratory and will be required to attend some courses in the biological sciences or other fields relevant to the study of biometry.

In addition to the grants in biometry, there are also fellowships under the National Defense Education Act available. Some appointments to graduate assistantships and research assistantships will also be made.

The Statistical Laboratory of the Catholic University of America is also expanding its activities into the areas of biomathematics and biometry. A training program and consulting service are being organized. Professor Edward Batschelet (on leave from the University of Basel, Switzerland) was appointed Visiting Professor in this program for the academic year 1960–61. Professor Harold Bergstrom of the Institute of Applied Mathematics of Chalmers Institutes of Technology (Goteborg, Sweden) was appointed Visiting Professor for the academic year 1960–61. He will primarily be engaged in research in probability theory. Professor D. Dugue of the Sorbonne (Paris, France) and Dozent T. E. Dalenius of Stockholm University are expected to visit Catholic University during the spring term 1961.

Requests for further information and application forms should be addressed to Professor Eugene Lukacs, Director, Statistical Laboratory, The Catholic University of America, Washington 17, D. C.

IOWA STATE UNIVERSITY

The National Science Foundation will sponsor a Summer Institute for College Teachers of Statistics at Iowa State University for the 11-week period from June 5 through August 18, 1961. The Departments of Statistics of three other universities, Kansas State, Utah State and the University of Wyoming, are cooperating with Iowa State's statistical center in presenting this institute.

Financial support in the form of stipends, dependency allowances and travel allowances will be awarded to 50 eligible applicants. All American college and university teachers who are, or who during the 1961–62 academic year will be, required to teach one or more courses in statistics as part of their regular assignments are eligible for consideration.

The institute is planned to provide additional basic training in statistics for present and prospective teachers who, though well-grounded in other fields. have limited backgrounds in statistics. Also it will provide more advanced courses and seminars designed to keep college and university teachers abreast of new developments.

Courses are scheduled in Statistical Methods, Theory of Statistics, Experimental Design, Survey Designs, Topics in Foundations of Probability and Statistics, and Intermediate Applied Decision Theory. In addition, an opportunity will be provided for those interested to observe a demonstration class in Principles of Statistics at the undergraduate level. The faculty will include the institute director, Dr. T. A. Bancroft, Director of the Iowa State University Statistical Laboratory and Head, Department of Statistics; Dr. R. J. Buehler, Associate Professor of Statistics, Iowa State University; Dr. H. T. David, Associate Professor of Statistics, Iowa State University; Dr. H. C. Fryer, Head of the Department of Statistics and Statistical Laboratory Director, Kansas State University; Dr. H. O. Hartley, Professor of Statistics, Iowa State University; the Institute Associate Director, Dr. D. V. Huntsberger, Associate Professor of Statistics, Iowa State University; and Dr. R. L. Hurst, Head of the Department of Applied Statistics and Statistical

Laboratory Director, Utah State University. Guest lecturers will present a series of special seminars.

Requests for information or application forms should be addressed to: The Director, Summer Institute in Statistics, 102 Service Building, Iowa State University, Ames, Iowa.

The Department of Statistics at Iowa State University will offer also eight applied courses in statistical theory and methods in its two 1961 summer sessions. These courses are planned primarily for graduate students or research workers with limited mathematical backgrounds who wish to use statistical techniques intelligently for application to other fields. In addition, a course on special topics in theoretical or applied statistics may be studied at the graduate level. Senior staff members will be available during most of the summer for consultations on research or special problems.

Students may register for either or both of the six-week summer sessions: June 5–July 12 and July 12–August 18. The complete list of statistics offerings for the first session is as follows: Stat. 401, "Statistical Methods for Research Workers "(at the level of Snedecor's *Statistical Methods*); Stat. 447, "Statistical Theory for Research Workers" (mainly theory of experimental statistics at the level of Anderson and Bancroft's *"Statistical Theory in Research"*; Stat. 411, "Experimetnal Designs for Research Workers," Stat. 599, "Special Topics;", Stat. 599A1, "Topics in Foundations of Probability and Statistics;" and Stat. 699, "Research." In the second session will be offered Stat. 402, a continuation of 401, Stat. 448, a continuation of 447; Stat. 421, "Survey Designs for Research Workers;" Stat. 599, Stat. 599A2, "Intermediate Applied Decision Theory (at the level of Blackwell and Girshick, *Theory of Games and Statistical Decisions*), and Stat. 699.

JOHNS HOPKINS UNIVERSITY

Beginning next September, the Department of Biostatistics at The Johns Hopkins University will offer an expanded program of study and research leading to Master of Science and Doctor of Science degrees. The curriculum has been modified by increasing the scope of the basic courses in statistical theory and statistical methods and by the addition of specialized courses including least squares and regression, stochastic processes, nonparametric methods, sampling and survey methods, biological assay, design of experiments and digital computer programming. These changes reflect a realization of the need for more intensively trained statisticians in the areas of biology, medicine and public health, and they are the direct result of the increasingly important role played by mathematics and statistics in all areas of scientific research.

The new program was made possible in part by the recent addition to the department of Drs. Allyn W. Kimball and David B. Duncan who together with Drs. Helen Abbey, Earl Diamond, John J. Gart and Margaret Martin form the permanent staff. In September, 1961, Dr. Norman T. J. Bailey of Oxford University will join the staff as visiting professor and will teach the course in stochastic processes. Additional appointments may be announced shortly.

The department has a limited number of liberal fellowships available, and interested students are invited to write to the Chairman, Department of Biostatistics, 615 North Wolfe Street, Baltimore 5, Md. for further information.

PURDUE UNIVERSITY

There will be three intensive courses in the general areas of statistical methods and quality control, design of experiments, and operations research at Purdue University this summer.

The course in Statistical Methods and Advanced Quality Control has been given at Purdue annually since 1947, being most recently revised in 1960. This course is designed for those who have had the equivalent of one of the intensive courses in statistical quality control given during and after the war, and who want to learn more about the statistical approach to industrial and research problems.

Topics to be studied during the ten-day quality control course are Significance Tests and Confidence Intervals, Significance of Differences, Linear Correlation and Regression, Single Sampling for Measurements, Sequential Sampling for Measurements, Multiple Correlation, and Analysis of Variance. Instructors include Profs. Irving W. Burr, the course director, and Charles R. Hicks, both of Purdue, Cecil C. Craig of the University of Michigan and Gayle McElrath of the University of Minnesota. The dates for this course are September 5–15.

The second course is on Design of Experiments and is for statisticians, quality control personnel, engineers, and others concerned with planning, analyzing, and interpreting the results of industrial experiments. The dates for this course are June 7–17. This will be the third year this course has been offered.

Professor Hicks, of the mathematical and statistical staff, the course director, emphasizes that this advanced course is for persons who have had previous statistical training, including work on tests of hypotheses, linear correlation, and at least an introduction to analysis of variance.

Topics included in the short course are Review of Analysis of Variance, Principles of Experimental Design, Variance Component Analysis, Randomized Blocks and Latin Squares, Factorial Experiments, Split Plot Designs, Confounding in Factorial Experiments, Incomplete Block Designs, Fractional Replications, and Introduction to Evolutionary Operations (EVOP).

In addition to Professor Hicks, instructors will include Prof. Clyde Y. Kramer, Virginia Polytechnic Institute, and Professor McElrath. Enrollment will be limited to a maximum of 25–30.

The third ten-day short course is on The Mathematical Techniques of Oper ations Research which will be offered at Purdue University for the second time this summer on June 5–15.

The course has been designed for statisticians, quality control analysis, engineers, and other technical personnel in industrial and management positions. Emphasis will be placed on the mathematical techniques of operations research and the application of these methods to current industrial and military problems.

These methods involve the construction of mathematical models representing the operation of industrial management or a military organization, and suggest the best solutions to problems involved. Among the topics to be discussed during the course are inventory control models, waiting line models, linear programming, simplex method, transportation methods, production scheduling models, search theory, cost-effectiveness studies, and systems analysis.

Instructors for the short course are Prof. Paul Randolph of Purdue, who is the short course director, Albert Madansky, of the RAND Corporation, and Prof. Bernard Lindgren, of the University of Minnesota. Enrollment in this course will be limited.

All three courses are sponsored jointly by the Statistical Laboratory and the Division of Adult Education at Purdue. Further information about any of the courses may be obtained by writing to the Division of Adult Education, Purdue University, Lafayette, Indiana.

SOUTHERN METHODIST UNIVERSITY

The National Science Foundation has awarded Southern Methodist University a grant of $26,400 to conduct a basic study of the feasibility of making the new Science Information Center at SMU a regional scientific and technical information center serving the needs of Southwestern industry and higher education.

Dr. William J. Graff, Jr., chairman of the SMU Department of Mechanical Engineering, is principal investigator for the study. He is assisted by Sam G. Whitten, SMU Science Librarian.

The building to house the new Science Information Center, a gift of local industrialists, is under construction and is scheduled for completion next August. A million dollar structure consisting of 80,000 sq. ft. on four floors, it will have space for half a million books. It will serve the SMU faculty and student body as well as the new Graduate Research Center. Space is provided in the building for administrative offices of the Graduate Research Center, whose director is Dr. Lloyd Berkner; for the SMU Map Library and Herbarium; and for the DeGolyer Geology Collection, in addition to the general document collections in science, mathematics, and engineering.

The study sponsored by the National Science Foundation is scheduled for completion in May, 1961. It will be based to a large extent on personal interviews and questionnaires aimed at the scientists and technicians who are engaged in research in the Southwest. An attempt will be made to find out what kind of information the research people need and in what subjects.

Also involved in the study are visits to and detailed studies of existing scientific information centers and an economic study of the area done by business economists.

The investigators expect to reach many research workers by questionnaires sent through local professional and scientific and technical societies, and will ask for employers aid in setting up personal interviews with research workers.

UNIVERSITY OF MINNESOTA

A special summer program of statistics in the Health Sciences will be held from June 13 to July 28, 1961 at the University of Minnesota. Courses in diverse areas of statistics at elementary and advanced levels will be given. Experts from around the nation will be teaching. Stipends are available to qualified students. For further information write to Biostatistics, 1226 Mayo, University of Minnesota, Minneapolis 14, Minnesota.

VIRGINIA POLYTECHNIC INSTITUTE

The 1961 session of the Southern Regional Graduate Summer Session in Statistics will be held at the Virginia Polytechnic Institute, Blacksburg, Virginia, from June 15 to July 22, 1961.

The Virginia Polytechnic Institute, Oklahoma State University, North Carolina State College, and the University of Florida have agreed to operate a continuing program of graduate summer sessions in statistics to be held at each institution in rotation. The program was instituted at Virginia Polytechnic Institute in the Summer of 1954.

The 1961 session, like previous sessions under this program, is intended to serve: 1) teachers of statistics and mathematics; 2) professional workers who want formal training in modern statistics; 3) research and engineering personnel who want intensive instruction in basic statistical concepts and modern statistical methodology; 4) Public Health statisticians who wish to keep informed about advanced specialized theory and methods; 5) prospective candidates for graduate degrees in statistics; and 6) graduate students in other fields who desire supporting work in statistics.

The session will last six weeks and courses will carry five quarter hours of credit. Not more than two courses may be taken for credit at any one session. The summer work in statistics may be applied as residence credit at any of the cooperating institutions, as well as certain other universities, in partial fulfillment of the requirements for a graduate degree. The program may be entered at any session, and consecutive courses will follow in successive summers so that it would be possible for a student to complete the course work in statistics for a Master's degree in three summers. Students must satisfy the remaining requirements for course work and thesis at the institution where they are to be admitted to candidacy. The advanced courses may be accepted as part of the Ph.D. program of the participating institutions.

A limited number of fellowships will be available for applicants from certain specialized areas. Doctoral courtesy will be honored for those holding Ph.D. or M.D. degrees.

The courses to be offered in statistics in 1961 at the Virginia Polytechnic Institute are as follows: Statistical Methods, Sampling Theory; Applied Statistics for Engineers and Physical Sciences; Theory I, Probability; Theory II, Statistical Inference; Theory III, Theory of Least Squares; Principles and Practices of High Speed Computing (enrollment limited to 14); Non-parametric Methods; and Multivariate Methods.

A number of courses in advanced mathematics will be available during the Summer Session. For a complete listing please consult the University timetable. A series of Colloquia involving recent developments in statistical theory and methods will be conducted during the special Summer Session.

Requests for application blanks for the summer session and for fellowships should be addressed to Dr. Boyd Harshbarger, Head, Department of Statistics, Virginia Polytechnic Institute, Blacksburg, Virginia.

CONTINENTAL CLASSROOM

Probability and Statistics, a nationally televised college-credit mathematics course, will be offered on the second semester of Continental Classroom beginning January 30, 1961.

Presented by Learning Resources Institute in cooperation with the Conference Board of the Mathematical Sciences, and telecast in color and black-and-white by the National Broadcasting Company, the course will be taught by Professor Frederick Mosteller, Chairman, Department of Statistics, Harvard University, and Professor Paul C. Clifford, Professor of Mathematics, Montclair (N. J.) State College.

More than 300 colleges and universities are expected to offer the televised course for college credit. It will be carried by 170 stations throughout the nation, from 6:30 to 7 a.m. Monday through Friday in each time zone.

NEW JOURNALS

Journal of the Forensic Science Society

Stuart S. Kind is the editor of a new journal, *The Journal of the Forensic Science Society* with editorial office c/o Rossett Holt, Pennal Ash Road, Harrogate, Yorkshire, England. Papers on biometry, especially in relation to human individuality, classification, and identification systems, will be welcomed.

Crop Science

Crop Science is the name of the new research journal which will appear in February 1961 as the official publication of the Crop Science Society of America, an affiliate of the American Society of Agronomy.

This new, bimonthly journal will carry research reports on breeding, genetics, physiology, ecology, and management of field crops, pastures, ranges, and turfgrasses from crop scientists in the U. S., Canada, and other countries.

Crop Science will be a companion publication to *Agronomy Journal*, the official organ of the ASA. It will alternate in publication dates with the *Journal* and will carry the articles (formerly publishable in the *Journal*) of direct interest to workers in the above-mentioned areas of research. *Agronomy Journal* will continue to carry the articles of wider agronomic scope—of interest to both crop and soil scientists, seed and weed technologists, plant pathologists, agricultural meteorologists, and others.

Publication of research reports in *Crop Science* will be open to members of the Crop Science Society—with joint membership in the American Society of Agronomy. Detailed information on publication in or subscription to *Crop Science* or *Agronomy Journal* or both may be obtained from the American Society of Agronomy, 2702 Monroe Street, Madison 5, Wisconsin.

TRAVEL GRANTS FOR ATTENDANCE AT THE INTERNATIONAL CONGRESS OF MATHEMATICIANS

Travel grants will be made to a number of mathematicians who wish to attend the International Congress of Mathematicians in Stockholm, on August 15–22, 1962. It is hoped that funds available through various sources may provide travel assistance for a considerable number of mathematicians.

There will be a greater effort than in the past to give aid to younger people. As grants will be made only to those who have filed applications, it is urgent that any who wish to receive a grant should fill out and file an application. Younger people are urged to file applications so that their cases can be considered. Applications can be obtained from the Division of Mathematics, National Academy of Sciences, National Research Council, Washington 25, D. C. by requesting an application for a travel grant to the 1962 International Congress.

The deadline for filing of applications is November 1, 1961, and an attempt will be made to announce the grants by January 1, 1962.

Awarding of grants will be made only to those persons whose applications have been received, in good order, by November 1. The selection will be made by a committee consisting of the regular Committee on Travel Grants of the Division of Mathematics of the National Academy of Sciences—National Research Council enlarged to include representatives of societies affiliated with the Division and representatives of various governmental agencies.

NEWS ABOUT MEMBERS

Laurence H. Baker is presently employed by Hy-Line Poultry Farms as a geneticist in Des Moines, Iowa.

David Bruce is Chief of Division Forest Fire Research, Pacific Northwest

Forest and Range Experiment Station, USDA. He formerly held the same position at the Southern Forest Experiment Station.

Martha W. Dicks is presently Assistant Professor of Human Nutrition at Montana State College.

Charles W. Dunnett was awarded a D.Sc. degree in statistics on completion of a two-year period of research at the University of Aberdeen, Scotland, under Dr. D. J. Finney, F.R.S. The title of his thesis was "The Statistical Theory of Drug Screening." He has now returned to his position as Head of the Statistical Design and Analysis Department at the Lederle Laboratories Division of the American Cyanamid Company, Pearl River, New York.

Lincoln J. Gerende is presently Instructor in Biostatistics, Department of Epidemiology and Public Health, Medical School, Yale University.

Leo A. Goodman is a Visiting Professor of Mathematical Statistics and Sociology at Columbia University during 1960-61 on leave from his position at the University of Chicago.

John Gurland, on leave from Iowa State University, is presently employed at the Mathematics Research Center of the University of Wisconsin in Madison, Wisconsin.

James A. Hagans of the Department of Preventive Medicine and Public Health of the University of Oklahoma has recently received his Ph.D. degree from that institution.

Carl E. Hopkins is now special consultant (Air Pollution Medical Studies) to the School of Public Health, UCLA.

Leo Katz has returned to his position as Head of the Department of Statistics, Michigan State University. He was on leave of absence serving as Scientific Liaison Officer for the Office of Navy Research in London, England, covering statistics and probability in Europe.

George H. Kennedy is presently employed by the Department of Health, Education and Welfare, U. S. Public Health Service in Bethesda, Maryland. He was formerly at the Biological Laboratories of the Department of Defense at Fort Detrick, Maryland.

John R. Kinzer has recently returned to his position of Professor of Psychology at Ohio State University, Columbia. He had been on a two-year leave of absence to work at the System Development Corporation, Santa Monica.

Gary F. Krause has given up his post as Instructor in Statistics at Kansas State University to become a full time graduate student in Statistics at the Virginia Polytechnic Institute at Blacksburg, Virginia.

Dr. Kuo Hua Lee has taken a position as an Associate Professor of Biostatistics in the University of Oregon Dental School, Portland.

William T. Lewish has taken a position as Special Service Engineer for the E. I. DuPont de Nemours and Co. in Wilmington, Delaware.

Alexander M. Mood formerly president of General Analysis Corporation became Vice-President of C-E-I-R, Inc. and Manager of the Los Angeles Research Center of C-E-I-R, when General Analysis Corporation merged with C-E-I-R.

Wesley L. Nicholson is on a four-month-special assignment as Professor of Mathematics at the University of Idaho. His permanent post is senior statistician for the General Electric Company in Richland, Washington.

L. M. Potter, formerly Research Associate in the Department of Poultry Science at the University of Connecticut, is now Associate Professor in the Department of Poultry Husbandry at the Virginia Polytechnic Institute at Blacksburg, Virginia.

Mr. J. N. K. Rao, an advanced graduate student in Statistics at Iowa State University, has been selected as the recipient of the George W. Snedecor Award in Statistics, which consists of a cash prize of $25.00, a year's membership in the Institute of Mathematical Statistics and a year's subscription to the *Annals of Mathematical Statistics*.

Maynard W. Shelly, II, formerly a psychologist in the Office of Naval Research is now a mathematician in the Logistics and Mathematical Statistics Branch of ONR.

John H. Smith is a Visiting Professor at the University of Chicago during the year. He is on sabbatical leave from the American University.

Jerry Warren is presently Assistant Professor in the Horticulture Department of the University of Nebraska. He was formerly with the Vegetable Crops Department of Cornell University.

The Biometric Society

BIOMETRICS

Founded by the Biometrics Section of the American Statistical Association

Table of Contents

Articles

Queries and Notes

Book Reviews

Number 2 June 1961 Volume 17

BIOMETRICS

Editor

Ralph A. Bradley

Editorial Borad

Former Editors

Gertrude M. Cox—*Founding Editor*

John W. Hopkins—*Past Editor*

Officers of the Biometric Society

General Officers

President: Leopold Martin

Secretary: M. J. R. Healy; *Treasurer*: M. A. Kastenbaum

Council: M. S. Bartlett, W. U. Behrens, C. I. Bliss, A. Buzzati-Traverso, D. G. Chapman, W. J. Dixon, C. W. Emmens, D. J. Finney, C. G. Fraga, Jr., Anne Lenger, C. C. Li, A. Linder, K. Mather, C. R. Rao, P. V. Sukhatme, G. Teissier, J. W. Tukey, G. S. Watson, F. Yates.

Regional Officers

Region	*President*	*Secretary*	*Treasurer*
Australasian	M. Belz	W. B. Hall	G. W. Rogerson
Belgian	M. Welsch	L. Martin	P. Gilbert
Brazilian	A. M. Penha	A. Conagin	A. Groszmann
British	J. A. Fraser Roberts	C. D. Kemp	P. A. Young
E. N. American	Oscar Kempthorne (Henry L. Lucas)	Erwin L. LeClerg	Donald A. Gardiner
French	Ph. L'Heritier	Sully Ledermann	Sully Ledermann
German	O. Heinisch	R. Wette	M. P. Geppert
Italian	G. Montalenti	R. Scossiroli	F. Sella
W. N. American	W. A. Taylor	W. A. Becker	Bernice Brown

National Secretaries

Denmark	N. F. Gjeddebæk	*Netherlands*	H. de Jonge
India	A. R. Roy	*Norway*	L. K. Strand
Japan	M. Hatamura	*Sweden*	H. A. O. Wold
	Switzerland	H. L. LeRoy	

ON ADDITIVITY IN THE ANALYSIS OF VARIANCE[1]

R. C. Elston

Department of Biostatistics, University of North Carolina
Chapel Hill, North Carolina, U. S. A.

INTRODUCTION

The problem relating to additivity in the analysis of variance is twofold. In the first place we wish to know whether we can remove any of the non-additivity present in our data, and in the second place we wish to know, given that it can be done, how to do so. Here two methods that have already been proposed for testing for non-additivity are developed from a somewhat different viewpoint, in an attempt to clarify their properties; and a generalization is given of a method that has been suggested for finding an appropriate transformation of the data.

TESTING FOR REMOVABLE NON-ADDITIVITY

Definition. If μ_{ij} is the subclass mean corresponding to the i-th level of A and to the j-th level of B in any 2-way classification, then we shall say that A and B are additive (or the μ_{ij} show additivity) if and only if there exist constants α_i , β_j such that

$$\mu_{ij} = \alpha_i + \beta_j , \qquad \text{all} \quad i, j.$$

For an n-way classification, denote the subclass means by μ_q , where q is an n-tuple of numbers designating the particular subclass; then there is additivity if and only if there exist constants α_{i_1} , β_{i_2} , $\cdots$, ν_{i_n} such that

$$\mu_q = \mu_{i_1 i_2 \cdots i_n} = \alpha_{i_1} + \beta_{i_2} + \cdots + \nu_{i_n} , \qquad \text{all} \quad i_1 , i_2 , \cdots , i_n .$$

Now whether or not the μ_q show additivity depends on the scale on which they are measured. In practice we measure our data on any scale that happens to be convenient, and from these measurements estimate the μ_q ; and, unless we transform the data, we are estimating the μ_q as measured on the same scale as that on which the data are measured. But, in many cases, the estimates we obtain are of far greater value to us if we can assume that the μ_q show additivity. Thus,

[1] Supported in part by departmental fellowship funds provided by U. S. Public Health Service.

BIOMETRICS

Officers of the Biometric Society

ON ADDITIVITY IN THE ANALYSIS OF VARIANCE[1]

R. C. Elston

Department of Biostatistics, University of North Carolina
Chapel Hill, North Carolina, U. S. A.

INTRODUCTION

The problem relating to additivity in the analysis of variance is twofold. In the first place we wish to know whether we can remove any of the non-additivity present in our data, and in the second place we wish to know, given that it can be done, how to do so. Here two methods that have already been proposed for testing for non-additivity are developed from a somewhat different viewpoint, in an attempt to clarify their properties; and a generalization is given of a method that has been suggested for finding an appropriate transformation of the data.

TESTING FOR REMOVABLE NON-ADDITIVITY

Definition. If μ_{ij} is the subclass mean corresponding to the i-th level of A and to the j-th level of B in any 2-way classification, then we shall say that A and B are additive (or the μ_{ij} show additivity) if and only if there exist constants α_i , β_j such that

$$\mu_{ij} = \alpha_i + \beta_j \ , \qquad \text{all} \quad i, j.$$

For an n-way classification, denote the subclass means by μ_q , where q is an n-tuple of numbers designating the particular subclass; then there is additivity if and only if there exist constants α_{i_1} , β_{i_2} , $\cdots$, ν_{i_n} such that

$$\mu_q = \mu_{i_1 i_2 \cdots i_n} = \alpha_{i_1} + \beta_{i_2} + \cdots + \nu_{i_n} \ , \qquad \text{all} \quad i_1 \ , i_2 \ , \cdots , i_n \ .$$

Now whether or not the μ_q show additivity depends on the scale on which they are measured. In practice we measure our data on any scale that happens to be convenient, and from these measurements estimate the μ_q ; and, unless we transform the data, we are estimating the μ_q as measured on the same scale as that on which the data are measured. But, in many cases, the estimates we obtain are of far greater value to us if we can assume that the μ_q show additivity. Thus,

[1] Supported in part by departmental fellowship funds provided by U. S. Public Health Service.

if we analyze a set of data y_q and find we have to reject the hypothesis that the corresponding μ_q show additivity, we should try and find a transformation of y_q to x_q such that the $E(x_q)$ show additivity (where E denotes expectation). Since the y_q are measured on a continuous scale, it is reasonable to consider only continuous functions of the y_q , and for such a transformation to be of any practical use it must be a *single-valued* function. With the additional reasonable requirement that the inverse function also be single-valued, this is equivalent to requiring the function (or its inverse) to be strictly monotonic.

From here on the symbol μ_q will be used to denote subclass means measured on a scale on which they show additivity, and we shall denote by y_q the data measured on any convenient scale. Suppose there exists a scale such that when the data are measured on it, and we denote these measurements by x_q ,

$$E(x_q) = \mu + \mu_q \quad \text{(where } \mu \text{ is any constant).} \tag{1}$$

Then the $E(x_q)$ show additivity. Furthermore, there is no loss of generality when we assume

$$\max \mu_q + \min \mu_q = 0, \qquad \max \mu_q \geq 0, \qquad \min \mu_q \leq 0. \tag{2}$$

(We assume μ_q lies within a definite range, determined by the extent of our data.)

If $E(y_q)$ is a linear function of $E(x_q)$, then the $E(y_q)$ show additivity. If $E(y_q)$ is a non-linear function of $E(x_q)$, then we can approximate this function by a quadratic polynomial, i.e.

$$E(y_q) \doteqdot a_0 + a_1E(x_q) + a_2[E(x_q)]^2. \tag{3}$$

If $E(y_q)$ is better approximated by such a quadratic polynomial than by any linear polynomial in $E(x_q)$, then, provided our data lie within the range where this quadratic polynomial is strictly monotonic, there exists a single-valued function such that, when applied to the data, the expectations of the transformed data come nearer to satisfying the condition for additivity.

Assuming exact equality in (3) and using (1) we may rewrite it as a polynomial in μ_q , say

$$E(y_q) = a + b\mu_q + c\mu_q^2 . \tag{4}$$

Differentiating with respect to μ_q we obtain $b + 2c\mu_q$, and here there is a maximum or minimum. Thus $E(y_q)$ will be strictly monotonic over the two ranges

$$\text{(i)} \qquad b + 2c \min \mu_q > 0$$

and

$$\text{(ii)} \qquad b + 2c \max \mu_q < 0.$$

Now using (2) it follows that if max μ_q − min $\mu_q < b/|c|$, (i) is satisfied; and if max μ_q − min $\mu_q < -b/|c|$, (ii) is satisfied (and in this case we must have $b < 0$).

It is clear from this that there is no point in considering the case $b = 0$, and so we can arbitrarily let $b = 1$ (i.e. divide our model through by b) and rewrite (4):

$$E(y_q) = \mu + \mu_q + c\mu_q^2 ,$$

where μ and c are appropriately redefined; and this is a strictly monotonic function over the range

$$\max \mu_q - \min \mu_q < \frac{1}{|c|}. \tag{5}$$

What we wish to do, then, is to assume a model of this form with the y_q normally and independently distributed, and test the null hypothesis that $c = 0$. If we reject this hypothesis, then, provided (5) holds, there exists a monotonic function of the data (whose inverse is a quadratic polynomial) that will bring us closer to additivity. In order to test this hypothesis we should like to find the reduction in sum of squares $R(\mathrm{I})$ due to fitting the model under the null hypothesis, and the reduction, $R(\mathrm{II})$, due to fitting the full model. These reductions are given by

$$R(\mathrm{I}) = \sum y_q^2 - \min \sum (y_q - \mu - \mu_q)^2,$$

and

$$R(\mathrm{II}) = \sum y_q^2 - \min \sum (y_q - \mu - \mu_q - c\mu_q^2)^2,$$

where the summations are over all observations, and μ, μ_q and c are estimated from the data. We can find $R(\mathrm{I})$ in the usual way; but when we set up the normal equations to find $R(\mathrm{II})$, we obtain a set of cubic equations that are not easy to solve.

The problem can be simplified, and an approximate solution obtained, as follows. For the sake of clarity consider the 2-way classification, one observation per subclass. Under the null hypothesis $c = 0$ we have

$$(\mathrm{I}) \quad E(y_{ij}) = \mu + \alpha_i + \beta_j ,$$

and otherwise

$$(\mathrm{II}) \quad E(y_{ij}) = \mu + \alpha_i + \beta_j + c(\alpha_i + \beta_j)^2.$$

Now the only extra term in the right hand side of (II) that depends on both i and j is the term $2c\alpha_i\beta_j$, and it is only if the absolute magnitude of this term is ever large that (II) can lead to a significantly smaller residual sum of squares than (I). We may therefore (neglecting the coefficient 2) consider the approximately equivalent model, (II′), given by

$$E(y_{ij}) = \mu + \alpha_i + \beta_j + c\alpha_i\beta_j \; ; \tag{6}$$

i.e. we assume $\alpha_i \fallingdotseq \alpha_i + c\alpha_i^2$ and $\beta_j \fallingdotseq \beta_j + c\beta_j^2$, and this is reasonable when we remember that c must satisfy (5) for monotonicity.

In order to obtain $R(\text{II}')$, the reduction in sum of squares due to fitting the model (6), it is necessary to minimize $\sum_{ij} (y_{ij} - \mu - \alpha_i - \beta_j - c\alpha_i\beta_j)^2$, which results in the following normal equations:

$$\begin{aligned}
\sum_{ij} (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{c}\hat{\alpha}_i\hat{\beta}_j) &= \sum_{ij} y_{ij} \\
\sum_{j} (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{c}\hat{\alpha}_i\hat{\beta}_j)(1 + \hat{c}\hat{\beta}_j) &= \sum_{j} y_{ij}(1 + \hat{c}\hat{\beta}_j) \\
\sum_{i} (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{c}\hat{\alpha}_i\hat{\beta}_j)(1 + \hat{c}\hat{\alpha}_i) &= \sum_{i} y_{ij}(1 + \hat{c}\hat{\alpha}_i) \\
\sum_{ij} (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{c}\hat{\alpha}_i\hat{\beta}_j)(\hat{\alpha}_i\hat{\beta}_j) &= \sum_{ij} y_{ij}\hat{\alpha}_i\hat{\beta}_j \; ,
\end{aligned} \tag{7}$$

where $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$ are a set of values for μ, α_i , β_j that minimize the above sum of squares. If we let $\sum_i \hat{\alpha}_i = \sum_j \hat{\beta}_j = 0$, we obtain

$$\sum_{ij} \hat{\mu} = \sum_{ij} y_{ij} \; , \tag{8}$$

$$\hat{\alpha}_i = \sum_j (y_{ij} - \hat{\mu} - \hat{\beta}_j)(1 + \hat{c}\hat{\beta}_j) / \sum_j (1 + \hat{c}\hat{\beta}_j)^2, \tag{9}$$

$$\hat{\beta}_j = \sum_i (y_{ij} - \hat{\mu} - \hat{\alpha}_i)(1 + \hat{c}\hat{\alpha}_i) / \sum_i (1 + \hat{c}\hat{\alpha}_i)^2, \tag{10}$$

$$\hat{c} = \sum_{ij} y_{ij}\hat{\alpha}_i\hat{\beta}_j / \sum_{ij} (\hat{\alpha}_i\hat{\beta}_j)^2. \tag{11}$$

We can obtain $\hat{\mu}$ directly from (8), but $\hat{\alpha}_i$, $\hat{\beta}_j$ and $\hat{c}$ have to be obtained by iteration. We first calculate $\hat{\alpha}_i$ and $\hat{\beta}_j$ from (9) and (10) assuming $\hat{c} = 0$, and then, using these values of $\hat{\alpha}_i$ and $\hat{\beta}_j$, we calculate ${}_1\hat{c}$, a first approximation to $\hat{c}$, from (11). Then assuming $\hat{c} = {}_1\hat{c}$ we recalculate $\hat{\alpha}_i$ and $\hat{\beta}_j$ from (9) and (10), and use these new values to obtain a second approximation to $\hat{c}$ from (11). The process is continued until all the estimates become stable. $R(\text{II}')$ is then given by

$$\begin{aligned}
&\sum_{ij} y_{ij}^2 - \sum_{ij} (y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{c}\hat{\alpha}_i\hat{\beta}_j)^2 \\
&\qquad = \hat{\mu} \sum_{ij} y_{ij} + \sum_i \hat{\alpha}_i \sum_j y_{ij} + \sum_j \hat{\beta}_j \sum_i y_{ij} + \hat{c} \sum_{ij} y_{ij}\hat{\alpha}_i\hat{\beta}_j
\end{aligned}$$

when $\sum_i \hat{\alpha}_i = \sum_j \hat{\beta}_j = 0$ and the estimates are such that this reduction is a maximum.

This method of solution was given by Ward and Dick [1952], but they considered the model (6) for a different reason; they wished to test the appropriateness of a multiplicative model of the form

$$E(y_{ij}) = (\mu + \alpha_i)(\mu' + \beta_j) = \mu\mu' + \mu'\alpha_i + \mu\beta_j + \alpha_i\beta_j \ ,$$

and this led them to assume the model (6). They state that the F-test for c is only approximate. If we assume as our full model

$$y_{ij} = \mu + \alpha_i + \beta_j + c\alpha_i\beta_j + \epsilon_{ij} \ ,$$

where the ϵ_{ij} are independent and $N(0, \sigma^2)$, then to test the null hypothesis that $c = 0$ we could take as our statistic

$$F = \frac{R(\mathrm{II}') - R(\mathrm{I})}{[\sum_{ij} y_{ij}^2 - R(\mathrm{II}')]/(n-1)} \ ,$$

where there would have been n degrees of freedom associated with the residual sum of squares if the model did not include the term $c\alpha_i\beta_j$. But whether or not this statistic actually does follow the F-distribution on 1 and $(n-1)$ degrees of freedom under the null hypothesis requires further investigation.

Now consider the sum of squares due to ${}_1\hat{c}$, the first approximation to $\hat{c}$, which is given by

$$\begin{aligned} R(\mathrm{II}'') - R(\mathrm{I}) &= \sum_{ij} y_{ij}^2 - \sum_{ij} (y_{ij} - \tilde{\mu} - \tilde{\alpha}_i - \tilde{\beta}_j - {}_1\hat{c}\tilde{\alpha}_i\tilde{\beta}_j)^2 \\ &\qquad - [\sum_{ij} y_{ij}^2 - \sum_{ij} (y_{ij} - \tilde{\mu} - \tilde{\alpha}_i\tilde{\beta}_j)^2] \\ &= -{}_1\hat{c} \sum_{ij} [{}_1\hat{c}(\tilde{\alpha}_i\tilde{\beta}_j)^2 - 2(\tilde{\alpha}_i\tilde{\beta}_j y_{ij} - \tilde{\mu}\tilde{\alpha}_i\tilde{\beta}_j - \tilde{\alpha}_i^2\tilde{\beta}_j - \tilde{\alpha}_i\tilde{\beta}_j^2)] \qquad (12) \end{aligned}$$

where $\tilde{\mu}$, $\tilde{\alpha}_i$ and $\tilde{\beta}_j$ are values of $\hat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$ satisfying (7), (9) and (10) when $\hat{c} = 0$, i.e. we have

$$y_{ij} = \tilde{\mu} + \tilde{\alpha}_i + \tilde{\beta}_j + \tilde{\epsilon}_{ij} \qquad (13)$$

where $\sum_{ij} \tilde{\epsilon}_{ij}^2$ is a minimum. Now from (11) we have ${}_1\hat{c} = \sum_{ij} y_{ij}\tilde{\alpha}_i\tilde{\beta}_j / \sum_{ij} (\tilde{\alpha}_i\tilde{\beta}_j)^2$. Using this and (13), and letting $\sum_i \tilde{\alpha}_i = \sum_j \tilde{\beta}_j = 0$, we find from (12),

$$R(\mathrm{II}'') - R(\mathrm{I}) = \frac{[\sum_{ij} \tilde{\alpha}_i\tilde{\beta}_j y_{ij}]^2}{\sum_{ij} (\tilde{\alpha}_i\tilde{\beta}_j)^2} = \frac{[\sum_{ij} \tilde{\alpha}_i\tilde{\beta}_j\tilde{\epsilon}_{ij}]^2}{\sum_{ij} (\tilde{\alpha}_i\tilde{\beta}_j)^2} .$$

From this it is seen that the sum of squares due to ${}_1\hat{c}$ is just the sum of squares proposed by Tukey [1949a] to test for non-additivity; the general formula given by Tukey [1955] is easily shown to be identical with the above for a two-way classification when $\sum_i \tilde{\alpha}_i = \sum_j \tilde{\beta}_j = 0$. Thus in Tukey's test for non-additivity we assume the model (II″)

$$y_{ij} = \mu + \alpha_i + \beta_j + c[\alpha_i\beta_j]_{ij} + \epsilon_{ij}\ ,$$

ϵ_{ij} independent, $N(0,\ \sigma^2)$, where $c[\alpha_i\beta_j]_{ij}$ indicates that the term is part of the interaction. This is to be interpreted as follows. We let

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

where $\sum_i (\alpha\beta)_{ij} = 0$ for all j and $\sum_j (\alpha\beta)_{ij} = 0$ for all i. This uniquely defines $\mu + \alpha_i + \beta_j$ as population parameters, but not α_i or β_j. Now for any set of α_i and β_j that are consistent with this definition let

$$\alpha_i\beta_j = \alpha_i' + \beta_j' + (\alpha\beta)_{ij}'$$

where $\sum_i (\alpha\beta)_{ij}' = 0$ for all j and $\sum_j (\alpha\beta)_{ij}' = 0$ for all i. Then this uniquely defines $(\alpha\beta)_{ij}'$, and this is what is represented above by $[\alpha_i\beta_j]_{ij}$. The parameter c in model (II″) can be thought of as the regression of the interaction effect $(\alpha\beta)_{ij}$ on the product of the main effects. We then test the null hypothesis that $c = 0$ using the statistic

$$F = \frac{R(\mathrm{II}'') - R(\mathrm{I})}{[\sum_{ij} y_{ij}^2 - R(\mathrm{II}'')]/(n-1)}. \tag{14}$$

Now since the y_{ij} are normally and independently distributed with variance σ^2, $R(\mathrm{II}'') - R(\mathrm{I})$ is distributed, for fixed $\tilde{\alpha}_i$ and $\tilde{\beta}_j$, as $\sigma^2\chi^2$ on 1 degree of freedom; and under the null hypothesis this is a central χ^2-distribution, for then $E(\tilde{\epsilon}_{ij}) = 0$. Also, since the distribution is the same for all fixed $\tilde{\alpha}_i$ and $\tilde{\beta}_j$, this is in fact the marginal distribution of $R(\mathrm{II}'') - R(\mathrm{I})$ as obtained from the joint distribution of $\tilde{\alpha}_i$, $\tilde{\beta}_j$ and $R(\mathrm{II}'') - R(\mathrm{I})$, as was pointed out by Tukey. Furthermore, $\sum_{ij} y_{ij}^2 - R(\mathrm{II}'') = \sum_{ij} \tilde{\epsilon}_{ij}^2 - [R(\mathrm{II}'') - R(\mathrm{I})]$ can be shown to be independently distributed as $\sigma^2\chi^2$ on $(n - 1)$ degrees of freedom, n being the number of degrees of freedom associated with $\sum_{ij} \tilde{\epsilon}_{ij}^2$. This shows that (14) follows an F-distribution on 1 and $(n - 1)$ degrees of freedom.[2]

Hamaker [1955] has given an example showing that the sum of squares due to $\hat{c}$ is more effective than the sum of squares due to ${}_1\hat{c}$ in accounting for non-additivity. But Tukey's test has the advantage of being computationally simpler and of resulting in a test statistic that is known to follow the F-distribution. It should be noted that for this test we have to assume a model that allows for no interaction effect other than can be accounted for by a term $c[\alpha_i\beta_j]_{ij}$. This is better

[2] A more complete proof of this is given on p. 132 of H. Scheffé's book "The Analysis of Variance" [Wiley, 1959]. Scheffé also gives another motivation for Tukey's test; this came to the author's notice after the present paper was written.

than assuming a model that allows for no interaction at all, but not as good as we might wish. If we have more than one observation per subclass, we can assume the more reasonable model

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} ,$$

$$\epsilon_{ijk} \text{ independent, } N(0, \sigma^2), \quad \text{and} \quad (\alpha\beta)_{ij} = c[\alpha_i\beta_j]_{ij} + \overline{(\alpha\beta)}_{ij}$$

and test the null hypothesis that $c = 0$, i.e. that

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \overline{(\alpha\beta)}_{ij} ,$$

$\overline{(\alpha\beta)}_{ij}$ being that part of the interaction effect that cannot be accounted for by a term $c[\alpha_i\beta_j]_{ij}$. Then we can use the test statistic (assuming that there are n degrees of freedom associated with the within subclass sum of squares)

$$F = \frac{R(\text{II}'') - R(\text{I})}{(\text{within subclass s.s.})/n} ;$$

this follows the F-distribution on 1 and n degrees of freedom under the null hypothesis, and, if condition (s) hold, provides an approximate test of whether the interaction sum of squares can be significantly reduced by a transformation of the data (using a monotonic function whose inverse is a quadratic polynomial).

The above can be easily generalized. For example, if we have a three-way classification, one observation per subclass, (6) becomes

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + c(\alpha_i\beta_j + \alpha_i\gamma_k + \beta_j\gamma_k),$$

using an obvious notation. The sum of squares due to ${}_1\hat{c}$ is then

$$R(\text{II}'') - R(\text{I}) = \frac{[\sum_{ijk} (\tilde{\alpha}_i\tilde{\beta}_j + \tilde{\alpha}_i\tilde{\gamma}_k + \tilde{\beta}_j\tilde{\gamma}_k)\tilde{\epsilon}_{ijk}]^2}{\sum_{ijk} (\tilde{\alpha}_i\tilde{\beta}_j + \tilde{\alpha}_i\tilde{\gamma}_k + \tilde{\beta}_j\tilde{\gamma}_k)^2} ,$$

provided $\sum_i \tilde{\alpha}_i = \sum_j \tilde{\beta}_j = \sum_k \tilde{\gamma}_k = 0$. If there is more than one observation per subclass, the summations are extended over all observations in each subclass.

When there are unequal subclass numbers, it is often inconvenient computationally to use the restriction that the class effects should sum to zero (this is especially the case when a large body of data is being analyzed with the aid of an electronic computer). Analogous to equation (11) we can find, for the two-way classification, the restriction-free formula

$$\begin{aligned} {}_1\hat{c} &= \frac{\sum (y_{ijk}[\tilde{\alpha}_i\tilde{\beta}_j]_{ij} - \tilde{\mu}[\tilde{\alpha}_i\tilde{\beta}_j]_{ij} - \tilde{\alpha}_i[\tilde{\alpha}_i\tilde{\beta}_j]_{ij} - \tilde{\beta}_j[\tilde{\alpha}_i\tilde{\beta}_j]_{ij})}{\sum [\tilde{\alpha}_i\tilde{\beta}_j]_{ij}^2} \\ &= \sum (y_{ijk}\tilde{\alpha}_i\tilde{\beta}_j - \tilde{\mu}\tilde{\alpha}_i\tilde{\beta}_j - \tilde{\alpha}_i^2\tilde{\beta}_j - \tilde{\alpha}_i\tilde{\beta}_j^2)/\sum [\tilde{\alpha}_i\tilde{\beta}_j]_{ij}^2 , \end{aligned}$$

the summations being over all observations. Substituting this value into the equation obtained from (12) on replacing $\tilde{\alpha}_i\tilde{\beta}_j$ by $[\tilde{\alpha}_i\tilde{\beta}_j]_{ij}$ (and allowing for more than one observation per subclass), we obtain

$$R(\text{II}'') - R(\text{I}) = \frac{[\sum (y_{ijk}\tilde{\alpha}_i\tilde{\beta}_j - \tilde{\mu}\tilde{\alpha}_i\tilde{\beta}_j - \tilde{\alpha}_i^2\tilde{\beta}_j - \tilde{\alpha}_i\tilde{\beta}_j^2)]^2}{\sum [\tilde{\alpha}_i\tilde{\beta}_j]_{ij}^2},$$

the summations again being over all observations. The denominator in this expression is that part of $\sum (\tilde{\alpha}_i\tilde{\beta}_j)^2$ that is due to an interaction effect; it is conveniently obtained in a way analogous to that by which the interaction sum of squares of the y_{ijk} is obtained, but with the variable $\tilde{\alpha}_i\tilde{\beta}_j$ replacing y_{ijk} for all k.

FINDING AN APPROPRIATE TRANSFORMATION FOR ADDITIVITY

Consider, as a function of μ_q, $E(y_q) = \mu + \mu_q + c\mu_q^2$. If $c > 0$, this is a convex function, while, if $c < 0$, this is a concave function. Also, the inverse function given by

$$\mu_q = \{-1 + \sqrt{1 - 4c[(\mu - E(y_q)]}\}/2c, \tag{15}$$

is conversely (where it is real-valued) concave if $c > 0$ and convex if $c < 0$.

Note that we can write the inverse function (15)

$$L(\mu_q) = [(E(y_q) + (1/4c) - \mu]^{1/2},$$

where $L(\mu_q)$ is a linear function of μ_q. This would suggest that, if we reject the hypothesis that $c = 0$, a transformation that might be expected to bring our data closer to additivity is the transformation $x = (y + k)^{1/2}$ where y is the observed value and k is estimated by $1/4\hat{c} - \hat{\mu}$. However, in view of the fact that the inverse of model (II′) (exemplified by equation (6) for the 2-way classification) is slightly different from (15), and that in any case such a model is unlikely to offer the best fit to the data, attention should not be restricted to this transformation alone. Tukey [1949a] suggests that we empirically seek out the best transformation in the class $x = (y + k)^p$, choosing $p < 1$ if ${}_1\hat{c} > 0$, and $p > 1$ [or trying $x = \log (y + k)$] if ${}_1\hat{c} < 0$. This advice regarding the choice of p follows from the fact that just as (15) is concave or convex according as $c > 0$ or $c < 0$, so $x = (y + k)^p$ is concave or convex according as $p < 1$ or $p > 1$. It is further suggested here that values of k approximating $(4 \mid {}_1\hat{c} \mid)^{-2p} - \hat{\mu}$ be tried first, since this is a solution to

$$[E(y_q) + (1/4{}_1\hat{c}) - \hat{\mu}]^{1/2} = [E(y_q) + k]^p,$$

when $E(y_q)$ is replaced by $\hat{\mu}$. However, it has been found empirically, from some artificially constructed examples, that this method of choosing k should only be used as a rough guide.

Now suppose we have two different transformations each of which we think may help in obtaining additivity. (For example, we might have two choices of p and k for the above general class of transformations; or we might have another entirely different transformation suggested by the nature of the data.) Then we might reasonably try and find the best linear combination of these two, i.e. a transformation of the form

$$x = d_1 f_1(y) + d_2 f_2(y),$$

where f_1 and f_2 are the two given functions and d_1 and d_2 are unknown coefficients. Tukey [1949b] has given a procedure for choosing d_1/d_2. His method can be extended to combine linearly more than two functions of the data, and the general method of obtaining suitable coefficients d_i will be given here.

Let $x = d_1 f_1(y) + d_2 f_2(y) + \cdots + d_p f_p(y)$. We will choose the p coefficients d_i given by the row vector $\mathbf{d}' = (d_1 , d_2 , \cdots , d_p)$ to maximize a certain ratio of sums of squares of the x's. Thus in a simple two-way classification Tukey considers maximizing

$$\frac{\text{row sum of squares of the } x\text{'s} + \text{column sum of squares of the } x\text{'s}}{\text{residual sum of squares of the } x\text{'s}}.$$

The method can be used to maximize any such ratio; and, provided we have a within-subclass sum of squares available, a general criterion would be to maximize

$$\frac{\text{within-subclass sum of squares of the } x\text{'s}}{\text{all interaction sums of squares of the } x\text{'s}},$$

since this would minimize any interaction variance component.

Let $s_1(f_i f_i)$ and $s_2(f_i f_i)$ be the numerator and denominator sum of squares respectively of the transformed data $f_i(y)$ in the ratio it is desired to maximize. Thus if we use the general criterion suggested above, $s_1(f_i f_i)$ is the within subclass sum of squares computed from the transformed data $f_i(y)$, and $s_2(f_i f_i)$ comprises all interaction sums of squares as computed after using the same transformation. Let $s_1(f_i f_j)$ and $s_2(f_i f_j)$ be respectively the corresponding numerator and denominator "sum of cross-products" of $f_i(y)$ and $f_j(y)$. For example, if $s_1(f_i f_i)$ were a row sum of squares, $s_1(f_i f_j)$ would be obtained as the sum of cross-products of row sums, divided by the usual divisor and less a corresponding correction term for the mean. If $s_1(f_i f_i)$ is the within subclass sum of squares, then $s_1(f_i f_j)$ is obtained exactly analogously, except that $f_i(y)$ and $f_j(y)$ are used to obtain cross-products instead of just $f_i(y)$ to obtain squares.

Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be $p \times p$ matrices whose elements in the i-th row and j-th column are $s_1(f_i f_j)$ and $s_2(f_i f_j)$ respectively.
Then

$$s_1(xx) = \sum_{i=1}^{p} d_i^2 s_1(f_i f_i) + 2 \sum_{i=1}^{p} \sum_{j>i} d_i \, d_j s_1(f_i f_j) = \mathbf{d}'\mathbf{S}_1 \, \mathbf{d},$$

where $s_1(xx)$ is the numerator sum of squares of the x's and $\mathbf{d}$ is a column vector, the transpose of $\mathbf{d}'$. Similarly

$$s_2(xx) = \mathbf{d}'\mathbf{S}_2 \, \mathbf{d}.$$

We thus wish to maximize $\mathbf{d}'\mathbf{S}_1\mathbf{d}/\mathbf{d}'\mathbf{S}_2\mathbf{d}$. Differentiating this with respect to $\mathbf{d}$ and equating to zero we obtain p equations that we may write in vector notation:

$$2 \frac{\mathbf{S}_1 \, \mathbf{d}}{\mathbf{d}'\mathbf{S}_2 \, \mathbf{d}} - 2 \frac{\mathbf{d}'\mathbf{S}_1 \, \mathbf{d}}{(\mathbf{d}'\mathbf{S}_2 \, \mathbf{d})^2} \mathbf{S}_2 \, \mathbf{d} = \mathbf{0},$$

where $\mathbf{0}$ is a column vector of p zeros; i.e.

$$\mathbf{S}_1 \, \mathbf{d} - \lambda \mathbf{S}_2 \, \mathbf{d} = \mathbf{0} \qquad (16)$$

where

$$\lambda = \frac{\mathbf{d}'\mathbf{S}_1 \, \mathbf{d}}{\mathbf{d}'\mathbf{S}_2 \, \mathbf{d}}.$$

Thus the largest value of λ is the value of the largest characteristic root of $(\mathbf{S}_1\mathbf{S}_2^{-1})$, and using this value of λ we choose $\mathbf{d}$ to satisfy (16). The individual d_i are not uniquely determined by (16), but all ratios d_i/d_j are. If we denote the element in the i-th row and j-th column of $(\mathbf{S}_1 - \lambda\mathbf{S}_2)$ by s_{ij}, we can easily find, for example when $p = 2$:

$$d_1/d_2 = -s_{12}/s_{11} \; ;$$

and when $p = 3$:

$$d_1/d_2 = (s_{13}s_{22} - s_{12}s_{23})/(s_{11}s_{23} - s_{12}s_{13}),$$

$$d_1/d_3 = (s_{12}s_{33} - s_{13}s_{23})/(s_{11}s_{23} - s_{12}s_{13}).$$

If we can find a $\mathbf{d}$ which makes all interaction sums of squares of the x's non-significant, we might assume the $E(x)$ are additive and so estimate our class effects on this scale. Unfortunately, however, it is not clear what hypotheses can be tested: any F-statistics we may obtain are influenced by our choice of $\mathbf{d}$. This procedure is therefore best restricted to the case where two independent sets of similar data are available; **d** can be estimated from one set of data, and using this **d** the other set of data can be transformed and analyzed. Then, if we assume normality of the transformed data, tests of significance will be valid; though only one set of the data should be used in making them.

The computational procedure in the case of unequal subclass numbers will not be discussed in any detail here. However it should be noted that a general solution has already been given, in another context, by Roy [1957]. If we equate Roy's $\mathbf{x}_r'$ to the vector of transformed variables associated with the r-th observation y_r, i.e. to $[f_1(y_r), f_2(y_r), \cdots, f_p(y_r)]$, then $\mathbf{S}_1$ and $\mathbf{S}_2$ in this paper can be equated to Roy's **S** and **S*** respectively.

REFERENCES

Hamaker, H. C. [1955]. Experimental design in industry. *Biometrics 11*, 257.

Roy, S. N. [1957]. *Some aspects of multivariate analysis*, 82-83. John Wiley and Sons, New York.

Tukey, J. W. [1949a]. One degree of freedom for non-additivity. *Biometrics 5*, 232.

Tukey, J. W. [1949b]. Dyadic anova, an analysis of variance for vectors. *Human Biology 21*, 65.

Tukey, J. W. [1955]. Answer to Query 113. *Biometrics 11*, 111.

Ward, A. C. and I. D. Dick [1952]. Non-additivity in randomized block designs and balanced incomplete block designs. *New Zealand Jour. Sci. and Tech. 33*, 430.

A NOTE ON SOME GROWTH PATTERNS IN A SIMPLE THEORETICAL ORGANISM

J. A. Nelder

National Vegetable Research Station, Wellesbourne, Warwick, England

INTRODUCTION

Hinshelwood and his co-workers [1946, 1955] have considered simple enzyme systems whose behaviour might mimic the growth of bacterial cultures under certain conditions. It is well known that the bacterial growth cycle of cell number on a given volume of medium divides itself into three main phases (Figure I). The first phase (or lag phase) represents a 'settling-in period' when growth may be slight or irregular; in the second phase, growth is logarithmic (that is log number is linearly related to time); finally in the third phase the relative growth rate falls from its constant value in the second phase to zero, often very rapidly, and the number in the colony becomes more or less steady. The lag time on transfer to a new culture may be generally defined by extending the linear portion of the logarithmic phase to the time axis (OT in Figure I). Considerable interest attaches to the form of the lag phase following (a) the transfer of cells from an ageing culture (i.e., one in the third phase) to a new medium and (b) the addition

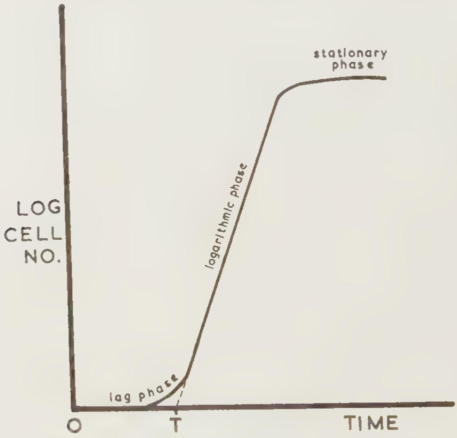


FIGURE I

Idealized Form of the Relation Between Log. Number of Cells and Time In A Bacterial Culture of a drug to the new medium. Hinshelwood has considered [1946, chap. IV *et seq.*] forms of lag produced in a very simple theoretical organism consisting of two linked enzymes, I and II, both of which are autocatalytic; thus enzyme I uses substrate outside the organism to produce more of itself and also an intermediate C, while II uses C to produce more of itself. Cell division is assumed to be geared to the amount of II. One of his equations of growth, however, appears to need amendment, and it is fortunate that with this amendment the equations have an explicit solution so that it becomes possible to replace his approximate treatment of the lag phase by an exact one in a number of different situations.

THE EQUATIONS OF GROWTH

We write X_1 and X_2 for the total amount of enzymes I and II in the system at any time, n for the total number of cells, Y for the total amount of intermediate, and $y = Y/n$ for the concentration of intermediate on a per cell basis. The equations for the growth of the enzymes are then given by

$$\frac{dX_1}{dt} = k_1X_1 \,, \tag{1}$$

$$\frac{dX_2}{dt} = k_2X_2y, \tag{2}$$

while the growth of the intermediate is given by

$$\frac{dY}{dt} = AX_1 - BX_2y - CY, \tag{3}$$

where AX_1 is the rate of production by enzyme I, $-BX_2y$ the rate of consumption by enzyme II, and CY the loss by diffusion and irrelevant chemical processes, which must be proportional to the concentration and the number of cells n. (The notation is the same as that in Dean and Hinshelwood [1955] except that we have used y for c and have introduced $Y = nc$ in their notation). Equation (3) differs essentially from the corresponding one given by Dean and Hinshelwood in having $dY/dt = d(nc)/dt$ instead of $n\ dc/dt$. The difference arises because their equation does not take into account the fall in concentration of the intermediate due to the expansion of the system. This fall will be given by $-(y/n)(dn/dt)$ in our notation, so that in terms of concentration the equation for y is given by

$$\frac{dy}{dt} = A\frac{X_1}{n} - B\frac{X_2}{n}y - Cy - \frac{y}{n}\frac{dn}{dt}$$

whence

$$n\frac{dy}{dt} + y\frac{dn}{dt} = \frac{d}{dt}(ny) = \frac{dY}{dt} = AX_1 - BX_2y - CY,$$

thus reducing to (3), as it must.

Hinshelwood postulates that the amount of enzyme II controls cell division which thus acts so as to keep X_2/n, the concentration of II, roughly constant. In what follows, we follow him in ignoring the changes in X_2/n and the diffusion coefficient C due to the discontinuous nature of the division process, and assume that the system can be taken as expanding continuously in such a way that X_2/n is constant, also that the area/volume ratio is constant so that C is unrelated to n. We therefore put $X_2 = n\theta$ where θ is a constant (and equal to β^{-1} in Dean and Hinshelwood's notation), whereupon equations (1), (2) and (3) become the linear set

$$\frac{dX_1}{dt} = k_1X_1\,, \qquad \frac{dX_2}{dt} = k_2\theta Y, \qquad \frac{dY}{dt} = AX_1 - Y(B\theta + C). \tag{4}$$

The general solution of (4) may be found by the usual methods and gives

$$\begin{aligned} X_1 &= Q_1k_1(\alpha + k_1)e^{k_1t}, \\ X_2 &= Q_1k_2\theta Ae^{k_1t} + Q_2 - Q_3k_2\theta e^{-\alpha t}, \\ Y &= Q_1k_1Ae^{k_1t} + Q_3\alpha e^{-\alpha t}, \end{aligned} \tag{5}$$

where Q_1, Q_2, and Q_3 are arbitrary constants and $\alpha = B\theta + C$. As t becomes large the organism tends to a steady state where X_1, X_2, and Y are all growing logarithmically in a constant ratio given by

$$X_1 : X_2 : Y = k_1(\alpha + k_1) : k_2\theta A : k_1A \tag{6}$$

When the organism is set going from a state different from the steady state given by (6), a lag phase will appear before the organism settles down into the logarithmic growth pattern. Two types of lag will be considered: (a) those produced by a decay in X_1 or Y (these possibilities are envisaged by Hinshelwood as being consequences of ageing in a culture) and (b) those produced by a drug which changes the value of one of the parameters in (4).

GROWTH FOLLOWING THE DECAY OF ENZYME AND INTERMEDIATE

(i) *The decay of X_1*

Suppose that at $t = 0$, $X_1 : X_2 : Y = pk_1(\alpha + k_1) : k_2\theta A : k_1A$, i.e. that relative to X_2 and Y, X_1 has declined to a fraction p of the steady state ratio given by (6). Solving for Q_1, Q_2 and Q_3 in (5) we find

$$n = X_2/\theta = \text{constant}$$
$$\times \left\{ \left(p/k_1 e \right)^{k_1 t} + (1 - p)\left(\frac{1}{k_1} + \frac{1}{\alpha} \right) - \frac{1 - p}{\alpha} e^{-\alpha t} \right\}. \qquad (7)$$

Without loss of generality we may define the unit of n so that $n = 1$ at $t = 0$ and the unit of t so that $k_1 = 1$. Then (7) becomes

$$n = pe^t + (1 - p) + \frac{1 - p}{\alpha}(1 - e^{-\alpha t}). \qquad (8)$$

As t becomes large

$$n \sim pe^t, \qquad \ln n \sim \ln p + t.$$

Hence the lag time T as defined above is obtained by putting $\ln n = 0$, whence $T = -\ln p$. Now α is a positive quantity in (8) and $(1 - e^{-\alpha t})/\alpha$ is a non-increasing function of α for any positive t; as $\alpha \to 0$, $(1 - e^{-\alpha t})/\alpha \to t$ while as $\alpha \to \infty$, $(1 - e^{-\alpha t})/\alpha \to 0$ so that extreme curves of the family (8) for fixed p are given by

$$n = pe^t + (1 - p)(1 + t),$$
and
$$n = pe^t + (1 - p). \qquad (10)$$

(ii) *The decay of Y*

If the amount of intermediate Y decays to a fraction p in the steady-state ratio (6) while the enzyme amounts remain fixed, then we obtain, corresponding to (8), the equation

$$n = e^t - [(1 - p)(1 - e^{-\alpha t})/\alpha].$$

The lag time is thus zero and the curves lie between

$$n = e^t - (1 - p)t, \qquad (\alpha = 0), \quad \text{and} \quad n = e^t, \qquad (\alpha = \infty).$$

GROWTH FOLLOWING CHANGES IN PARAMETER VALUES

One possible explanation of drug action is that it changes values of the parameters in the growth equations; thus an intermediate may be inactivated by the drug and be rendered more or less unavailable to the next enzyme in the chain, or active sites on an enzyme may be blocked, and so on. In the simple organism considered here growth depends on 5 parameters, k_1, k_2, θ, A, and $\alpha = B\theta + C$. We shall not consider variations in θ or k_1; if θ varied, then the drug would change the mean cell size and so alter C as well, while if k_1 is permanently changed, the final steady-state growth rate will be changed. We thus consider effects on k_2, A, and α which do not change the mean cell size or affect the steady-state growth rate after adaptation. In addition the drug will be assumed to be present in excess so that the effect on the parameters can be taken as fixed and independent of time. The results are summarised below:

(iii) *Fall in A*

The drug affects the production of intermediate by enzyme I. If A is reduced to a fraction p of its former value and cells are in their logarithmic phase before treatment then, following the same conventions as before

$$n = pe^t + (1 - p) + [(1 - p)(1 - e^{-\alpha t})/\alpha]$$

so that the effect is identical to that of a corresponding fall in the amount of X_1 (equation 8).

(iv) *Fall in k_2*

Here the drug affects the autocatalysis of enzyme II, and

$$n = pe^t + (1 - p),$$

which is an extreme case, as $\alpha \to \infty$, of equation (8). The lag in both these cases is given by $T = -\ln p$.

(v) *Rise in α*

In this case the drug is assumed to mop up the intermediate as it is produced, thus increasing C and hence α. The quantity α can also be increased by increasing B, the rate of use of the intermediate by enzyme II, but this seems a less likely effect of drug action. The cell number is given by

$$n = \frac{1}{1 + p\alpha}\left[(1 + \alpha)e^t - \alpha(1 - p) - \frac{(1 - p)}{p}(1 - e^{-p\alpha t})\right] \qquad (11)$$

where $p > 1$, since α is assumed to have increased. The lag is given by $T = \ln[(1 + p\alpha)/(1 + \alpha)]$. The extreme members of (11) are given

$$n = e^t \quad \text{for} \quad \alpha = 0, \qquad (12)$$

and

$$n = \pi e^t + (1 - \pi) \quad \text{for} \quad \alpha = \infty, \qquad \pi = 1/p. \qquad (13)$$

Equation (12) of course gives logarithmic growth with no lag, while (13) has been met previously as one of equations (10).

GROWTH FOLLOWING DISTURBANCE OF TWO QUANTITIES OR PARAMETERS

In this section we consider only the growth following the disappearance of intermediate accompanied by changes in X_1 or one of the parameters A, k_2, or α. Three of these cases produce the same equation for n, namely those having a fall in X_1, A, or k_2 to a fraction p of their steady state value. All these give

$$n = pe^t + (1 - p) - [p(1 - e^{-\alpha t})/\alpha],$$

with lag $T = -\ln p$, and extreme curves

$$n = pe^t + (1 - p) - pt \quad \text{for} \quad \alpha = 0,$$

$$n = pe^t + (1 - p) \qquad \text{for} \quad \alpha = \infty.$$

The effect of α is not large here; the curves having $\alpha = 0$ have a slightly more sudden rise to the logarithmic asymptote.

For a rise in α to $p\alpha$ accompanied by an initial value of $Y = 0$ we get

$$n = \frac{1}{1 + p\alpha}\left[(\alpha + 1)e^t + (p - 1)\alpha - \frac{\alpha + 1}{p\alpha}(1 - e^{-p\alpha t})\right].$$

The lag is $T = \ln [(1 + p\alpha)/(1 + \alpha)]$ and the extreme curves are given by

$$n = e^t - t \quad \text{for} \quad \alpha = 0,$$

$$n = \pi e^t + (1 - \pi), \qquad \pi = 1/p \quad \text{for} \quad \alpha = \infty.$$

EXAMPLES OF GROWTH CURVES

Cases (i) and (iii) give the same family of curves. Examples of these are shown in Figure II (a), (b), and (c) for the extreme values of α (0 and ∞), and for $\alpha = 1$, the curves being given for lags equal to 2, 4, 8, and 16 generation times. (On our standard scale with $k_1 = 1$, the generation time is $\ln 2 = 0.693$ units.) Case (iv) is covered by Figure II(b), while case (v) has as extremes the simple zero-lag line $\ln n = t$ for $\alpha = 0$ and the curves of Figure II(b) for $\alpha = \infty$. For $\alpha = 1$, curves very similar to those of Figure II(b) are obtained. In Figure II(d) an example is given for case (ii) with the extreme values $\alpha = 0$, $p = 0$; for $p < 1$, these curves are sigmoid and asymptotic to the no-lag line $\ln n = t$ from below, the slope at the origin being p for all finite α.

For growth following complete loss of Y accompanied by a change in X_1, A, k_2 or α, the curves are similar to those of Figure II(b) for all positive α, and examples are not given.

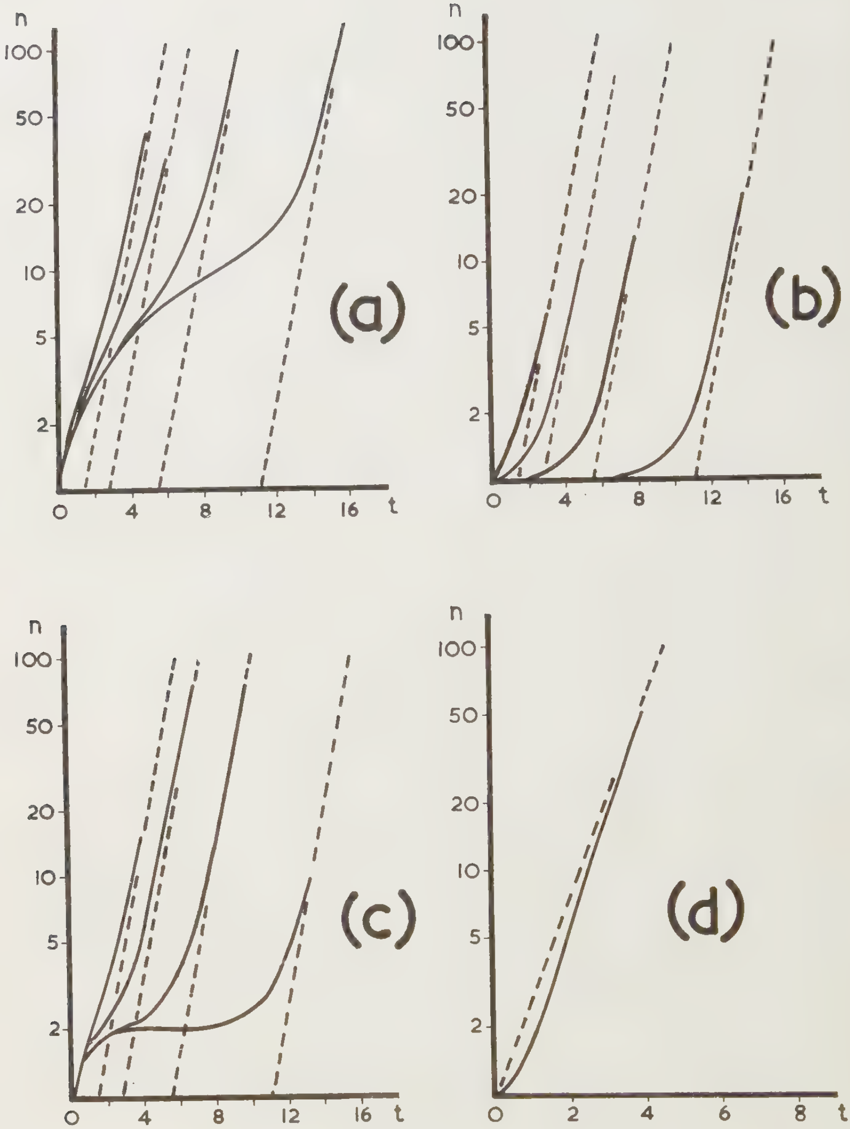


FIGURE II

EXAMPLES OF GROWTH CURVES AFTER DISTURBANCE FROM THE STEADY STATE

(a) Curves of the family $n = pe^t + (1 - p)(1 + t)$
(b) Curves of the family $n = pe^t + (1 - p)$
(c) Curves of the family $n = pe^t + (1 - p)(2 - e^{-t})$
(d) The curve $n = e^t - t$

Notes: the ordinate is $\ln n$ throughout. In (a), (b), and (c) p takes the values 2^{-2}, 2^{-4}, 2^{-8}, and 2^{-16}.

DISCUSSION

Though the "organism" discussed in this note is a very simple one, the curves show that it is capable of a variety of different growth patterns after disturbance from the steady state. The curves exemplified in Figure II(c) are somewhat similar to one of those shown in Figure 10 of Hinshelwood's book (Hinshelwood [1946]), and described by him as an "irregular growth curve". However we have not succeeded in producing a curve such as that shown in Figure 11 of the same work where growth is very slight for about 4 generation times and the logarithmic phase is then entered very rapidly indeed. When the lag is as short as 4 generation times, none of the curves produced by the model has a near-horizontal early part. It is only when the lag reaches about 10 generation times that we find in some cases a substantial period at the beginning when growth is very slight.

One way of producing a curve of the type of Hinshelwood's Figure 11 might be to postulate a drug which combined with the intermediate and prevented its use by enzyme II, the drug itself becoming inactivated by the combination. In this situation the amount of intermediate available to enzyme II would be kept very low until all the drug had been immobilized after which it would rise to the stable state level. The general question of the behaviour of the organism 'growing away' from a drug after immobilizing it is complex and does not seem amenable to exact solution. Similarly the growth equations for Hinshelwood's other theoretical organisms, such as the one having two cyclically linked enzymes (Hinshelwood [1946], p. 81 *et seq.*), or alternative pathways (*ibid.* p. 150 *et seq.*) appear to have no explicit general solution, though of course numerical solutions for given values of the constants could easily be obtained.

A striking feature of the model is the occurrence of similar or identical growth equations of n following different changes of the steady state. Since these changes have quite distinct biological interpretations, it is clear that it will rarely be possible to deduce from the shape of a growth curve the sort of mechanism likely to have produced it.

SUMMARY

The growth equations for a simple organism proposed by Hinshelwood, and consisting of two linked autocatalytic enzymes, are shown to produce a linear system under certain conditions.

Solution of these equations gives the form of the lag in growth following (i) decay of one enzyme and/or the intermediate and (ii) changes in the values of the parameters such as might be brought about by drug action.

Examples of the types of curves produced are given, and their resemblance to actual bacterial growth curves discussed. The production of identical curves by several distinct mechanisms in the model is pointed out.

REFERENCES

Hinshelwood, Sir Cyril [1946]. *The Chemical Kinetics of the Bacterial Cell.* Oxford University Press, Oxford, England.

Dean, A. C. R., and Hinshelwood, Sir Cyril [1955]. Reaction Patterns of a Coliform 'Organism. *Progressive Biophysics & Biophys. Chem.* 5, 1–40.

THE PARTIAL DIALLEL CROSS[1]

O. Kempthorne
Iowa State University, Ames, Iowa, U. S. A.
AND
R. N. Curnow[2]
Agricultural Research Council Unit of Statistics,
University of Aberdeen, Aberdeen, Scotland.

1. *Introduction*

The diallel cross, which is composed of all possible single crosses among a group of inbred lines, is now a common plan of investigation in both plant and animal breeding. Its modern use starts apparently with the development of the concepts of general and specific combining ability by Sprague and Tatum [1942]. The diallel cross is used to estimate the genetic components of the variation among the yields of the crosses (see, for instance, Hayman [1954a, b]; Jinks and Hayman [1953]; Griffing [1956a, b] and Kempthorne [1956, 1957]). It is also used to estimate the actual yielding capacities of the crosses. This information may be employed, for example, to select the best four inbred lines from which to develop a four-way cross. We shall discuss only the relatively simple situation in which there are no maternal effects, i.e. reciprocal crosses are identical and so need not be made, and in which there is no interest in the performance of the inbred lines themselves. The important questions concerned with replication over years and locations will not be considered. The methods used and conclusions reached in this paper may be capable of extension to more complicated situations.

With no reciprocal crosses or crosses resulting from selfing or crossing within the same inbred line, there are ${}^nC_2 = n(n - 1)/2$ possible single crosses among n inbred lines. This number of possible crosses increases rapidly with n. When $n = 6$ it is only 15, when $n = 20$ it is 190, and when $n = 50$ it is 1,225. With facilities available for testing only a limited number of crosses, a diallel cross may only be possible between a relatively small number of inbred lines. If only a small number of inbred lines are tested, the estimate of the variance of the

[1] Journal Paper No. J-3909 of the Iowa Agricultural and Home Economics Experiment Station, Ames, Iowa. Project 890.

[2] The work on this paper was completed while the latter author was visiting Iowa State University on a Harkness Fellowship of the Commonwealth Fund.

general combining abilities among the whole population of potentially available inbred lines will be subject to a large sampling error and many potentially high yielding inbred lines left completely untested. The question arises, therefore, of whether a design involving more inbred lines but only a sample of all possible crosses between them may not be preferable. A particular method of sampling the diallel cross will be considered and its efficiency in estimating the genetic variance components, the differences between the yields of the various crosses and the general combining abilities of the lines discussed.

2. *The Partial Diallel Cross*

The best method to be used in deciding how to sample the diallel cross would be to specify, in general terms, the probabilities of sampling each particular cross and each particular pair of crosses and then choose these probabilities so as best to satisfy the aims of the experiment. We have not been able to do this. We shall consider only the particular method of sampling the crosses suggested by G. W. Brown in about 1948. The method certainly achieves some balance and whether a much better one exists is rather doubtful. It has already been used twice at Iowa State University (Jensen, [1959]; Sprague, unpublished.).

Assume that the breeder can handle a total of $ns/2$ crosses where n is the number of inbred lines and s is a whole number greater than or equal to 2. Clearly, n and s cannot both be odd. The n inbred lines are numbered at random from 1 to n and the following crosses sampled:—

$$\begin{array}{l}
\text{line } 1 \times \text{lines } k+1, k+2, \cdots, k+s \\
\text{line } 2 \times \text{lines } k+2, k+3, \cdots, k+1+s \\
\quad \cdots \qquad \cdots \qquad \cdots \\
\text{line } i \times \text{lines } k+i, k+i+1, \cdots, k+i-1+s \\
\quad \cdots \qquad \cdots \qquad \cdots \\
\text{line } n \times \text{lines } k+n, k+n+1, \cdots, k+n-1+s,
\end{array}$$

where $k = (n + 1 - s)/2$, and is a whole number, and all the numbers above n are to be reduced by multiples of n so as to be between 1 and n. For k to be a whole number, either n is odd and s even or n is even and s odd. Each line occurs in s crosses and the total number of crosses sampled is $ns/2$. $s = n - 1$ corresponds to the complete diallel cross. The reader may have noticed the analogy between this method of sampling the diallel cross and the experimental design for blocks of two plots in which the s blocks containing treatment number i also contain treatments numbered $k + i, k + i + 1, \cdots, k + i - 1 + s$. These circulant designs are discussed by Kempthorne [1953]. To every design for blocks of two plots in which each treatment occurs the same number of times, there corresponds a method of sampling the diallel cross in which each line is involved in the same number of crosses, and conversely. Treatments i and j occurring in the same block corresponds to sampling the cross $i \times j$. Any balanced incomplete block design corresponds to the complete diallel cross. Partial diallel crosses corresponding to other two plot per block designs have not been considered. Partial diallel crosses corresponding to other circulant designs, including some with n and s both even, (Zoellner and Kempthorne, [1954]), could be studied by methods very similar to those of this paper.

We shall assume that the yield from the cross $i \times j$ in replicate l can be written

$$y_{ijl} = \mu + r_l + g_i + g_j + s_{ij} + e_{ijl} ,$$

where μ is a general effect, r_l a replicate effect, g_i and g_j are the parental effects (sometimes called general combining abilities), s_{ij} is the effect of the non-additivity of the parental effects (sometimes called specific combining ability) and e_{ijl} is the plot error. We shall further assume that g_i , s_{ij} and e_{ijl} are independently normally distributed with zero means and variances σ_g^2 , σ_s^2 and σ_e^2 . The motivation for this model is as follows. The yield of cross ij in replicate l is assumed to consist of three parts combining additively: an effect of the cross, a replicate effect and an error due to plot deviation and also to segregation within the cross, if there is any. As regards the cross effect, different progeny of the same cross are full sibs, and progeny of two different crosses involving a common line are half sibs, and this is the only genetical structure. If the lines are a random sample from a large population, the cross effect can be represented by

$$g_i + g_j + s_{ij}$$

in which g_i , g_j and s_{ij} are assumed to be independent random variables with variances σ_g^2 , σ_g^2 and σ_s^2 respectively, where

$$\sigma_g^2 = \text{Cov (H.S.)}$$

and

$$\sigma_s^2 = \text{Cov (F.S.)} - 2\,\text{Cov (H.S.)},$$

H.S. denoting half-sibs and F.S. denoting full-sibs.

The diallel cross can be used with multi-flowered plants and, if the plants used are a random sample of the population of plants, the above covariances are covariances of relatives in that population. If furthermore that population has random mating structure, is not inbred and there is no linkage,

$$\text{Cov (H.S.)} = \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{16}\sigma_{AA}^2 + \cdots$$

and

$$\text{Cov (F.S.)} = \tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_D^2 + \tfrac{1}{4}\sigma_{AA}^2 + \cdots .$$

If the parents are individuals arising by inbreeding to degree F in such a population, then (Cockerham [1954], Kempthorne [1957]·

$$\text{Cov (H.S.)} = \left(\frac{1+F}{4}\right)\sigma_A^2 + \left(\frac{1+F}{4}\right)^2 \sigma_{AA}^2 + \cdots$$

and

$$\text{Cov (F.S.)} = \left(\frac{1+F}{2}\right)\sigma_A^2 + \left(\frac{1+F}{2}\right)^2 \sigma_D^2 + \left(\frac{1+F}{2}\right)^2 \sigma_{AA}^2 + \cdots .$$

If F equals unity, i.e. the parents are completely inbred, Cov (H.S.) is equal to the covariance of parent and offspring in the original random mating population and Cov (F.S.) is the genotypic variance in the original population. If the lines are only partially inbred, genetic interpretation is possible only if the degree of inbreeding is constant over the lines and is known.

For a general value of F, all we know about σ_g^2 and σ_s^2 is that the former, σ_g^2 = Cov (H.S.), involves only the variances due to additive effects and interactions of additive effects and that the latter, σ_s^2 = Cov (F.S.) $-$ 2 Cov (H.S.), does not involve the variance due to additive effects, σ_A^2 . If the lines are completely inbred,

$$\sigma_g^2 = \tfrac{1}{2}\sigma_A^2 + \tfrac{1}{4}\sigma_{AA}^2 + \cdots$$

and

$$\sigma_s^2 = \sigma_D^2 + \tfrac{1}{2}\sigma_{AA}^2 + \cdots .$$

The total genotypic variance is $2\sigma_g^2 + \sigma_s^2$ and, if epistacy can be ignored, the additive variance is $2\sigma_g^2$. The ratio of additive to total genotypic variance would then be

$$2\sigma_g^2/(2\sigma_g^2 + \sigma_s^2),$$

and the square root of twice the ratio of dominance to additive variance, which, if gene frequencies are equal to $\frac{1}{2}$, is interpretable as the average degree of dominance, is

$$\sqrt{\sigma_s^2/\sigma_g^2} .$$

If the lines are not inbred at all, i.e. $F = 0$,

$$\sigma_g^2 = \tfrac{1}{4}\sigma_A^2 + \tfrac{1}{16}\sigma_{AA}^2 + \cdots$$

and

$$\sigma_s^2 = \tfrac{1}{4}\sigma_D^2 + \tfrac{1}{8}\sigma_{AA}^2 + \cdots .$$

If epistacy can be ignored, the ratio of additive to total genotypic variance is

$$\sigma_g^2/(\sigma_g^2 + \sigma_s^2),$$

and the average degree of dominance, (see above)

$$\sqrt{2\sigma_s^2 \ \sigma_g^2} .$$

The plot error variance σ_e^2 will contain a component due to genetic variability within crosses. In fact it will contain the total genotypic variance less the covariance of full-sibs and this is zero only if F equals unity. The environmental variability in the plot error e_{ijl} consists of plot to plot variability plus environmental variability particular to each individual and, if there were competition, would also contain a component from this force.

The above model seems entirely appropriate for the consideration of the estimation of σ_g^2 and σ_s^2 , which are known as half the variance of general combining ability and the variance of specific combining ability respectively. Likewise it seems entirely appropriate for the estimation of the covariances of half-sibs and full-sibs, or estimation of related quantities like the average degree of dominance.

It is less appropriate for the consideration of the yields of possible multi-way crosses, such as two-way or four-way crosses, because the plant breeder will be in the position of having a specified set of lines and is interested in the potentialities within that set of lines rather than within a random set of lines that he might have obtained. But it appears to us that even in this case the plant breeder has no alternative to assuming an additive model in which the yield of cross $i \times j$ is made up of an effect due to line i, an effect due to line j, and an effect due to the non-additivity of the effects of lines i and j. Earlier an attempt was made to consider the situation in the framework that there was a finite population of N lines from which n were drawn at random and the partial diallel cross made among these n lines. Unfortunately, considerable mathematical difficulties were encountered and no neat result obtained. We shall therefore look at the problem of estimating all possible single crosses by the use of the simple additive model described above.

3. *Estimation of the Variance Components*

We shall estimate the general combining abilities $g_1, \cdots, g_n$ by least squares, i.e. by choosing $\hat{\mu}, \hat{g}_1, \cdots, \hat{g}_n$ to minimize

$$\sum_s (y_{ij} - \hat{\mu} - \hat{g}_i - \hat{g}_j)^2,$$

where $\sum_s$ denotes summation over the sampled crosses and y_{ij} the mean yield of cross $i \times j$. With the imposed constraint

$$\sum_{i=1}^{n} \hat{g}_i = 0,$$

this leads to the equation $\hat{\mu} = \bar{y}$, the average of the y_{ij}'s, and to the following equations for $\hat{g}_1, \cdots, \hat{g}_n$,

$$\sum_{j=1}^{n} a_{ij}\hat{g}_j = \sum_{r=0}^{s-1} \left[y_{i,i+k+r} - \frac{2G}{ns} \right] = Q_i, \qquad (i = 1, 2, \cdots, n), \tag{3.1}$$

where G is the grand total of cross averages and $a_{ii} = s$, all i, and $a_{ij} = a_{ji} = 1$ if cross $i \times j$ is sampled and $a_{ij} = a_{ji} = 0$ otherwise. The $n \times n$ matrix $\mathbf{A} = [a_{ij}]$ is a symmetric circulant matrix and therefore so is its inverse, $\mathbf{A}^{-1} = [a^{ij}]$. (The general form of a circulant matrix is given in the appendix at the end of this paper. A symmetric circulant matrix has the additional property that the element in the ith row and jth column is the same as the element in the jth row and ith column for all i and j.) Being a symmetric circulant, the elements a^{ij} of the matrix $\mathbf{A}^{-1}$ are functions only of $| i - j |$, the positive value of $(i - j)$. We shall therefore write $a^{ij} = a^{|i-j|}$. By multiplying together rows of $\mathbf{A}$ with columns of $\mathbf{A}^{-1}$,

$$\sum_{r=k}^{k+s-1} a^r = 1 - sa^0 \tag{3.2}$$

and

$$\sum_{r=k}^{k+s-1} a^{t+r} = -sa^t, \qquad t = 1, \cdots, n-1. \tag{3.3}$$

In (3.3), the $t + r$ index in the summation is to be reduced by multiples of n so as always to be between 0 and $n - 1$. By summing (3.3) from $t = 1$ to $t = n - 1$, and using (3.2), the sum of any row of $\mathbf{A}^{-1}$ is

$$\sum_{r=0}^{n-1} a^r = 1/2s. \tag{3.4}$$

The total sum of squares of cross averages in the analysis of variance of the partial diallel cross can be decomposed as follows:

$$\sum_s (y_{ij} - \bar{y})^2 = \sum_{i=1}^{n} \hat{g}_i Q_i + \sum_s (y_{ij} - \bar{y} - \hat{g}_i - \hat{g}_j)^2,$$

where $\bar{y}$ is the mean of the cross averages.

Now from (3.1), and (3.4)

$$\sum_{i=1}^{n} \hat{g}_i Q_i = \sum_{i=1}^{n} \sum_{j=1}^{n} a^{ij} Q_i Q_j$$

$$= \sum_{i,j=1}^{n} \sum_{l,m=0}^{s-1} a^{ij} y_{i,i+k+l} y_{j,j+k+m} - 2G^2/ns. \qquad (3.5)$$

Fixing a particular sampled cross, $i \times (i + k + l)$, and considering all sampled crosses, $j \times (j + k + m)$, related to it, either as full or half sibs, and using (3.2), (3.3) and (3.4),

$$E\left(\sum_{j=1}^{n} \sum_{m=0}^{s-1} a^{ij} y_{i,i+k+l} y_{j,j+k+m}\right) = \frac{\mu^2}{2} + (a^0 + a^{k+1})$$

$$\cdot\left\{\frac{\sigma_e^2}{r} + \text{Cov (F.S.)} - 2\,\text{Cov (H.S.)}\right\} + \text{Cov (H.S.)}. \qquad (3.6)$$

Summing (3.6) over i and l, again using (3.2), (3.3) and (3.4), and substituting an expression for $E(G^2)$, (3.5) gives

$$E\left(\sum_{i=1}^{n} \hat{g}_i Q_i\right) = (n - 1)\left\{\frac{\sigma_e^2}{r} + \text{Cov (F.S.)} - 2\,\text{Cov (H.S.)}\right\}$$

$$+ s(n - 2)\,\text{Cov (H.S.)}$$

Table 1 shows the analysis of variance of the partial diallel cross, with the expected values of the mean squares written in terms of σ_g^2 and σ_s^2 instead of in terms of Cov (F.S.) and Cov (H.S.).

TABLE 1

Analysis of Variance of Partial Diallel Cross

Source	d.f.	Expected values of mean squares
Replicates	$r - 1$	
General Combining Ability	$n - 1$	$\sigma_e^2 + r\sigma_s^2 + \{rs(n - 2)/(n - 1)\}\,\sigma_g^2$
Specific Combining Ability	$n(s/2 - 1)$	$\sigma_e^2 + r\sigma_s^2$
Replicates × Crosses	$(r - 1)(ns/2 - 1)$	σ_e^2
Total	$rns/2 - 1$	

With a complete diallel cross there are always many more degrees of freedom for the s.c.a. (specific combining ability) mean square than for the g.c.a. (general combining ability) mean square. Unless σ_g^2 is small compared to $\sigma_e^2 + r\sigma_s^2$, this may be a serious disadvantage in the estimation of σ_g^2. The partial diallel cross does allow the $(ns/2 - 1)$ degrees of freedom for crosses to be split more evenly between the two mean squares. Clearly, for σ_s^2 to be estimable, s has to be greater than 2. With $s = 3$, there will be nearly twice as many degrees of freedom for the g.c.a. mean square as for the s.c.a. mean square. With $s = 4$, the degrees of freedom will be approximately equal. Table 2 shows the breakdown of the degrees of freedom for an experiment involving 2 replicates of 120 plots each when the crosses are (i) a complete diallel cross ($s = n - 1 = 15$), (ii) a partial diallel cross with $s = 3(n = 80)$ and (iii) a partial diallel cross with $s = 4(n = 60)$. The determination of the optimal numbers of replicates r, inbred lines n, and crosses per inbred line $s/2$ will involve the unknown values of σ_g^2/σ_e^2 and σ_s^2/σ_e^2 as well as the cost function of the experiment. Another difficulty in deciding on a value for s is that an exact specification of the aims of the experiment is necessary. One possible aim of the experiment is to estimate components of genetic variance such as the additive and dominance portions and the goodness of the design will depend on these and on the inbreeding coefficient of the lines. A compromise may have to be made between various aims. The proportion of the genetic variance that is additive (Fisher, [1918]) and the average degree of dominance (Comstock and Robinson, [1948, 1952]) are both, under severe assumptions (see Section 2), functions of σ_s^2/σ_g^2. Again, measures

TABLE 2

Degrees of Freedom for Complete and for Partial Diallel Cross Using 2 Replicates of 120 Plots Each

Source	Degrees of freedom		
	Complete diallel cross	Partial diallel cross	
		$s = 3$	$s = 4$
Replicates	1	1	1
G. c. a.	15	79	59
S. c. a.	104	40	60
Replicates × Crosses	119	119	119
Total	239	239	239

of heritability will be functions of σ_g^2 , σ_s^2 , and σ_e^2 . A design that minimizes the variance of the estimates of σ_g^2 and σ_s^2 does not necessarily minimize the variance of the estimates of these functions. For example, by increasing the degrees of freedom for the g.c.a. mean square at the expense of the degrees of freedom for the s.c.a. mean square, the variance of the estimate of σ_s^2 is increased and the covariance between the estimates of σ_g^2 and σ_s^2 made more negative. This may result in an increase in the variance of the estimate of σ_s^2/σ_g^2 , despite the decreases in the variance of the estimate of σ_g^2 .

4. *Comparison of Partial Diallel Cross with Other Designs for Estimating the Variance Components*

A comparison of the partial or complete diallel cross with other designs for estimating the variance components is complicated by all the factors mentioned at the end of the last section. The experiment analogous to what Comstock and Robinson [1952] called Experiment I would be to cross each of m randomly chosen inbred lines used as "sires" with m different sets of f randomly chosen inbred lines used as "dams". This type of experiment has been widely used in poultry to estimate covariance of full-sibs and half-sibs. No inbred line would be used both as a "sire" and a "dam". The analogy is not complete since the experimental material considered by Comstock and Robinson was produced from random matings among plants of the F_2 generation of a cross of two inbred lines, whereas we are considering the progeny resulting directly from crosses among a set of inbred lines. The analysis of variance of Experiment I is shown in Table 3. The variance of the estimates of both σ_g^2 and σ_s^2 with this design can be shown to be greater than the corresponding variances for the partial diallel cross with $n = m$ and $s = 2f$. The two designs involve the same number of

TABLE 3

ANALYSIS OF VARIANCE OF AN EXPERIMENT IN WHICH m INBRED LINES ARE EACH CROSSED TO A DIFFERENT SET OF f INBRED LINES

Source	d.f.	Expected values of mean squares
Replicates	$r - 1$	
"Sires"	$m - 1$	$\sigma_e^2 + r\sigma_s^2 + r(f + 1)\,\sigma_g^2$
"Dams in Sires"	$m(f - 1)$	$\sigma_e^2 + r\sigma_s^2 + r\sigma_g^2$
Replicates × Crosses	$(r - 1)(mf - 1)$	σ_e^2
Total	$rmf - 1$	

crosses but the partial diallel cross uses f fewer inbred lines. The covariance between the estimates of σ_g^2 and σ_s^2 is less negative with the diallel cross than with Experiment I. The experiment analogous to Experiment II of Comstock and Robinson [1952] would be to make all the crosses between m randomly chosen inbred lines used as "sires" and f randomly chosen inbred lines used as "dams". No inbred line would be used both as a "sire" and a "dam". The analysis of variance of this design is shown in Table 4. An estimation procedure that does not use both the sire and dam mean squares to estimate σ_g^2 can be shown to be inferior to the estimation procedure possible with a partial diallel cross with $n = m$ and $s = 2f$, i.e. a partial diallel cross involving the same number of crosses but f fewer inbred lines. Comstock and Robinson [1952] have suggested that, when possible, Experiment II should be designed with $m = f$. This is not necessarily optimal, but it does permit a simple pooling of the dam and sire mean squares. Unfortunately, it also results in many more degrees of freedom for the s.c.a. mean square than for the g.c.a. mean square. There will be $(m - 1)/2$ times as many degrees of freedom for the s.c.a. mean square as for the g.c.a. mean square. Unless σ_g^2 is small compared with $\sigma_e^2 + r\sigma_s^2$, this may be a serious disadvantage in the estimation of σ_g^2. Indeed one of the main general reasons for using the partial diallel cross is that it gives a reasonable number of degrees of freedom for the g.c.a. mean square.

A comparison of the partial diallel cross with Experiment III of Comstock and Robinson [1952] and other more complicated designs will not be attempted. Experiment III may be superior to the partial diallel cross. A complete comparison of the two designs would again involve the values of unknown parameters and would also raise questions

TABLE 4

Analysis of Variance of All Crosses Between Two Different Sets of m and f Inbred Lines

Source	d.f.	Expected values of mean squares
Replicates	$r - 1$	
"Sires"	$m - 1$	$\sigma_e^2 + r\sigma_s^2 + rf\sigma_g^2$
"Dams"	$f - 1$	$\sigma_e^2 + r\sigma_s^2 + rm\sigma_g^2$
"Sires × Dams"	$(m - 1)(f - 1)$	$\sigma_e^2 + r\sigma_s^2$
Replicates × Crosses	$(r - 1)(mf - 1)$	σ_e^2
Total	$rmf - 1$	

regarding the sensitivity of the estimation procedures to departures from some of the genetic assumptions.

All the designs discussed so far have involved crosses between lines both of which were drawn at random from the population of inbred lines. One result of this is that all of the variation between crosses can be used to estimate either σ_g^2 or σ_s^2 . There are two serious disadvantages involved in the use of common testers that are not a random sample from the same population as the inbred lines. Firstly, the general combining ability of a line has to be defined as the average performance of that line when crossed to all members of a certain population. The average performance of an inbred line when crossed to a set of common testers will almost certainly not be the same as its average performance when crossed to all other members of the population of inbred lines. This could seriously invalidate the estimate of σ_g^2 . Secondly, there would seem to be little or no reason to assume that the variance of the specific combining abilities among crosses between inbred lines and common testers is equal to the variance of the specific combining abilities among crosses between the inbred lines themselves. These considerations strongly suggest that common testers should not be used to estimate the values of σ_g^2 and σ_s^2 for the population of inbred lines, unless the common testers are themselves a random sample from that same population.

5. *Comparing the Yielding Capacities of the Crosses*

We shall make the same assumptions and use the same notation as in the previous sections of this paper. We shall consider the partial diallel cross as a method of estimating the yielding capacities of all the possible single crosses among n inbred lines. Inbred lines are often developed and then crossed in an attempt to produce crosses with a high yielding capacity. The yielding capacity of all crosses in the diallel cross and all sampled crosses in the partial diallel cross can be estimated in each of two ways. Firstly, they can be estimated by their mean yields in the experiment. Secondly, specific combining ability can be ignored and the yielding capacity of the cross $i \times j$ estimated by $\hat{\mu} + \hat{g}_i + \hat{g}_j$.[3] Clearly, the latter method has to be used to estimate the yielding capacity of all the unsampled crosses in the partial diallel cross.

We shall designate the methods of estimation A and B, where A and B are defined as follows:

[3] The two estimates could possibly be given appropriate weights and combined into a single estimate. For the partial diallel cross these weights are rather involved and vary from cross to cross. In this paper, the two methods of estimation will be considered separately and then compared.

A. Unsampled crosses estimated by $\hat{\mu} + \hat{g}_i + \hat{g}_j$ but sampled crosses estimated by cross means y_{ij} .

B. Both unsampled and sampled crosses estimated by $\hat{\mu} + \hat{g}_i + \hat{g}_j$.

A more precise statement will be made later, but one may summarize briefly that B will be preferable to A only if $r\sigma_s^2/\sigma_e^2$ is small.

A third method C of estimating the yielding capacities of the crosses is to cross all n inbred lines with each of t common testers. As mentioned previously, general combining ability is not a property of the line alone but a property of the line relative to the population of lines to which it has been crossed. To speak of the general combining ability of a line without reference to that population is vague if not meaningless. If common testers are to be used to estimate the yielding capacities of crosses between the inbred lines then we have to make the assumption, and it is a big assumption, that the general combining ability of each inbred line relative to the common testors is the same as its general combining ability relative to all the other inbred lines. We hope that the importance of this assumption will not be lost in the discussion that follows. Without it there is no basis for comparing the diallel cross with common testers. In method C, we shall estimate the yielding capacity of cross $i \times j$ by $\hat{\mu} + \hat{g}_i + \hat{g}_j$ where $\hat{g}_i$ and $\hat{g}_j$ are the estimates of the general combining abilities of lines i and j with the common testers. Like B, C will be preferable to A only if $r\sigma_s^2/\sigma_e^2$ is small. Neither of the two designs alternative to the diallel cross considered in Section 4 can be used to estimate the differences between crosses without confounding some of these differences with other contrasts between the general combining abilities. They will therefore not be considered further.

In all three methods, A, B and C, the errors involved in comparing any two crosses will be composed of two parts. There will be errors arising from an inadequate estimation of specific combining abilities and from the specific combining abilities that enter into the estimates $\hat{g}_i$. There will also be errors arising from the plot effects. To measure the error in comparing any two crosses, we shall calculate the expected value of the square of this error. It will be composed of two parts, one involving σ_s^2 and the other σ_e^2 . The criterion Q, used to measure the efficiency of the various methods, will be the average, over all ${}^{n(n-1)/2}C_2$ possible comparisons among the $n(n - 1)/2$ crosses, of this expected square error. The values of Q for methods A, B and C will be written Q_A , Q_B and Q_C respectively.

With all three methods of estimation, comparisons can only be made between crosses involving the inbred lines actually tested. This is an obvious statement, but that any design not involving each inbred line at least once has $Q = \infty$ should be realized. As a result, the criterion Q cannot be used to decide on the optimal division of resources between the number of inbred lines on the one hand and the number of replicates and crosses per inbred line on the other. The number of inbred lines tested n is fixed and we need to estimate the yielding capacities of all crosses among them. A decision on the value of n would presumably involve a compromise between the accurate assessment of a relatively small number of crosses and the relatively inaccurate assessment of a large number of crosses. The more crosses tested the more intense can be the selection applied to them but the larger will be the errors involved in that selection. Problems of this kind are discussed in a series of papers by Finney [1958 a, b; 1960]. The methods he used to study mass selection in plant breeding cannot be applied directly to the present problem because of the unequal accuracies and the correlations involved in the estimation of the different crosses and because of the genetic covariance σ_g^2 between crosses involving a common line. Also, the aim may not be simply the selection of the best crosses but, for example, the selection of the best four inbred lines from which to form a four-way cross.

Two other limitations of the present approach should be noted. Firstly, we are considering only one property of the crosses, namely "yield". With more than one property, a compromise may have to be made and a design chosen that is optimal for one quantity but not for another. Secondly, we are considering all crosses among a set of inbred lines. Of more interest sometimes would be all crosses between two different sets of inbred lines that are chosen to bring together certain properties that are present in one set but not in the other.

The algebra needed to calculate the values of Q for the two methods of estimation A and B is very tedious. The comparisons among the crosses have to be divided into the three categories: sampled versus sampled, unsampled versus unsampled, and sampled versus unsampled. Using equations (3.2), (3.3) and (3.4) to sum over these categories, we find,

$$Q_A = \frac{2}{s(n-2)(n+1)} [\{2n(n-2)sa^0 + ns^2 - 2ns - n + 2\} \cdot \{\sigma_s^2 + (\sigma_e^2/r)\} + \{n^2 - 2ns - n + 2\}s\sigma_s^2]$$

and

$$Q_B = \frac{2}{s(n-2)(n+1)} [\{2n(n-2)sa^0 - n + 2\}\{\sigma_s^2 + (\sigma_e^2/r)\} + \{n^2 - 5n + 2\}s\sigma_s^2],$$

where a^0 is, as before, the diagonal element of the inverse matrix $\mathbf{A}^{-1}$. The value of Q for C, the common tester method, is

$$Q_C = \frac{4(n-1)}{(n+1)t}\,[\sigma^2_{s(t)} + (\sigma^2_e/r)] + 2\sigma^2_s\,,$$

where t is the number of common testers, r the number of replicates, and $\sigma^2_{s(t)}$ the variance of specific combining ability among the crosses between the inbred lines and the common testers. The quantity $2\sigma^2_s$ occurs in Q_C because the estimate of each cross comparison $i \times j$ versus $l \times m$ is biased by the amount $s_{ij} - s_{lm}$. To compare A, B and C for the same total number of crosses and experimental plots, we shall assume that $t = s/2$, and that the two values of r are the same. The assumption that $\sigma^2_{s(t)} = \sigma^2_s$ has already been criticized in connection with the estimation of σ^2_g and σ^2_s. However an assumption has to be made about the ratio $\sigma^2_{s(t)}/\sigma^2_s$ and $\sigma^2_{s(t)}/\sigma^2_s = 1$ seems the most reasonable. Other values could be studied by using the general form for Q_C given above.

A comparison between the two methods A and B is immediately possible.

$$Q_A - Q_B = \frac{2n(s-2)}{(n-2)(n+1)}\,[(\sigma^2_e/r) - \sigma^2_s], \quad \text{for all} \quad n \quad \text{and} \quad s.$$

As is otherwise obvious, the two methods are identical when $s = 2$. If $s > 2$, A is preferred to B, i.e. $Q_A < Q_B$, if and only if $\sigma^2_e/r\sigma^2_s < 1$. If $r > 1$, a decision between A and B could be based on an estimate of $\sigma^2_e/r\sigma^2_s$ obtained from the analysis of variance of Table 1. To compare method C with methods A and B, the values of a^0 are required. The matrix $\mathbf{A}^{-1}$ can be easily determined when $s = 2$ and when $s = n - 1$ (the complete diallel cross). When $s = 2$, $\mathbf{A}^{-1}$ is generated as a circulant by the first row

$$\left(\frac{n}{4}, \frac{n}{4} - 1, \frac{n}{4} - 2, \cdots, -\frac{n}{4} + \frac{3}{2}, -\frac{n}{4} + \frac{1}{2}, -\frac{n}{4} + \frac{1}{2}, -\frac{n}{4} + \frac{3}{2}, \cdots, \frac{n}{4} - 1\right)$$

and, when $s = n - 1$, $\mathbf{A}^{-1}$ has all diagonal elements equal to $(2n - 3)/2(n - 1)(n - 2)$ and all non-diagonal elements equal to $-1/2(n - 1)(n - 2)$. Substituting $a^0 = n/4$ in Q_A and Q_B, we have that, when $s = 2$,

$$Q_A = Q_B = Q_C + \frac{(n-1)}{(n+1)(n-2)}\,[(n-2)(n-3)\{\sigma^2_s + (\sigma^2_e/r)\} - 8\sigma^2_s].$$

When $n = 3$, the complete diallel cross is preferred to common testers unless $\sigma_s^2 = 0$, when they are equally efficient. When $n = 5$, the partial diallel cross with $s = 2$ is preferred to using one common tester if and only if $\sigma_e^2/r\sigma_s^2 < \frac{1}{3}$. When $n \geq 7$, using one common tester is to be preferred to the partial diallel cross with $s = 2$ for all values of $\sigma_{e.}^2/r\sigma_s^2$. Substituting $a^0 = (2n - 3)/2(n - 1)(n - 2)$ in the expression for Q_B, we have that for the complete diallel cross $(s = n - 1)$

$$Q_B + \frac{4}{(n - 2)(n + 1)} \{(n - 3)(\sigma_e^2/r) + (3n - 5)\sigma_s^2\}.$$

The complete diallel cross with estimation method B is therefore always to be preferred to using $(n - 1)/2$ common testers. The complete dialled cross with estimation method A will be even better if $\sigma_{e.}^2/r\sigma_s^2 < 1$.

Values of a^0 for s intermediate to $s = 2$ and $s = n - 1$ are not so easy to obtain. An explicit expression for a^0 is derived in an appendix at the end of this paper. Expressions for the other elements of the inverse matrix $\mathbf{A}^{-1}$ are also derived and may be of use in obtaining variances for comparing the yielding capacities of particular crosses. The formula for a^0 is

$$a^0 = \frac{1}{n}\left\{\frac{1}{2s} + \sum_{j=1}^{n-1} \frac{\sin\dfrac{j\pi}{n}}{s\sin\dfrac{j\pi}{n} - \sin\dfrac{(n - s)j\pi}{n}}\right\}.$$

Values for a^0 calculated from this formula on an IBM 650 electronic computer agreed very well with known values for $s = n - 1$. The agreement for $s = 2$ and n large was not so good. The values of a^0 for $s = 2$ in Table 5 were calculated by $a^0 = n/4$. All other values were calculated electronically using the formula above. They are thought to be sufficiently accurate for our purpose but not necessarily accurate to the number of decimals shown.

Table 6 shows the values of Q_A, Q_B and Q_C for these same values of n and s. The first figure for each method is the coefficient in Q of σ_e^2/r and the second is the coefficient of σ_s^2. The bordered squares are those in which the common tester method is to be preferred to the diallel cross method. They are the squares for which s/n is small. In all other squares, the best method is A or B according as $\sigma_e^2/r\sigma_s^2$ is less than or greater than $\sigma_{e.}^2/r\sigma_s^2 = 1$. The only exception to this is that when $s = 2$, A and B are equivalent. Also, for each value of n there are, presumably, intermediate values of s for which the decision between the diallel cross and common testers depends on the unknown value of $\sigma_e^2/r\sigma_s^2$. The examples of this in Table 6 are $n = 101$, $s = 10$; $n = 50$, $s = 7$;

TABLE 5

VALUES OF a^0 FOR VARIOUS n AND s

n	s									
	2	3	4	5	6	7	10	20	50	100
101	25.25		2.588		0.8332		0 2369	0.06515	0.02059	0.01005
100		6.031		1.365		0.5530				
75	18.75		1.969		0.6523		0.1978	0.06027	0.02034	
74		4.515		1.044		0.4387				
51	12.75		1.391		0.4836		0.1616	0.05575	0.02020	
50		3.119		0.7514		0.3326				
21	5.25		0.6739		0.2732		0.1162	0.05132		
20		1.394		0.3868		0.2007				
11	2.75		0.4248		0.2014		0.1056			
10		0.7689		0.2604		0.1589				
9	2.25		0.3721		0.1904					
8		0.6458		0.2387		0.1548				
7	1.75		0.3233		0.1833					
6		0.5139		0.2250						
5	1.25		0.2917							
4		0.4167								
3	0.75									

$n = 75$, $s = 74$; $a^0 = 0.01361$; $n = 9$, $s = 8$; $a^0 = 0.1339$.

$n = 10$, $s = 3$; $n = 8$, $s = 3$ and $n = 5$, $s = 2$, when the partial diallel cross is preferred to common testers if and only if $\sigma_e^2/r\sigma_s^2 < 0.25$, 0.42, 0.63, 5.93 and 0.33 respectively.

For given values of n and σ_e^2/σ_s^2, Table 6 could be used to compare the values of Q_A, Q_B and Q_C for different values of r and s. r will have to be greater than 1 if σ_e^2 is to be estimated and s greater than 2 if σ_s^2 is to be estimated.

6. *Estimation of General Combining Abilities*

The variance of the estimate of the difference between two general combining abilities using the partial diallel cross is

TABLE 6

COEFFICIENTS* OF σ_e^2/r AND σ_s^2 IN Q_A, Q_B AND Q_C FOR A RANGE OF VALUES OF n AND s

n		s																			
		2		3		4		5		6		7		10		20		50		100	
	A	100.0	101.9	23.9	25.8	10.28	12.13	5.46	7.26	3.38	5.14	2.29	4.01	1.16	2.70	0.62	1.82	1.04	1.04	2.00	0.00
101	B	100.0	101.9	23.9	25.8	10.25	12.17	5.40	7.32	3.30	5.22	2.19	4.11	0.94	2.86	0.26	2.18	0.081	2.00	0.040	1.96
100	C	3.92	5.92	2.61	4.61	1.96	3.96	1.57	3.57	1.31	3.31	1.12	3.12	0.78	2.78	0.39	2.39	0.157	2.16	0.078	2.08
	A	74.0	75.9	17.8	19.7	7.82	9.59	4.20	5.92	2.69	4.36	1.86	3.48	0.99	2.46	0.72	1.64	1.38	0.68	$s = 74$ 2.00	0.00
75	B	74.0	75.9	17.8	19.7	7.77	9.66	4.11	6.01	2.57	4.46	1.73	3.62	0.78	2.67	0.24	2.13	0.080	1.97	0.053	1.95
74	C	3.89	5.89	2.60	4.60	1.95	3.95	1.56	.56	1.30	3.30	1.11	3.11	0.78	2.78	0.39	2.39	0.156	2.16	0.105	2.10
	A	50.0	51.8	12.3	14.0	5.53	7.21	3.06	4.66	2.05	3.57	1.50	2.93	0.95	2.15	0.94	1.34	2.00	0.00		
51	B	50.0	51.8	12.2	14.1	5.45	7.29	2.94	4.78	1.89	3.73	1.30	3.14	0.63	2.47	0.22	2.06	0.078	1.92		
50	C	3.85	5.85	2.56	4.56	1.92	3.92	1.54	3.54	1.28	3.28	1.10	3.10	0.77	2.77	0.38	2.39	0.15	2.15		
	A	20.0	21.6	5.38	6.77	2.75	3.97	1.77	2.73	1.43	2.24	1.28	1.82	1.24	1.25	2.00	0.00				
21	B	20.0	21.6	5.28	6.88	2.55	4.17	1.45	3.05	1.03	2.65	0.75	2.35	0.43	2.05	0.19	1.81				
20	C	3.64	5.64	2.41	4.41	1.82	3.84	1.45	3.45	1.21	3.21	1.03	3.03	0.73	2.75	0.36	2.36				
	A	10.0	11.3	2.96	3.69	1.92	2.37	1.59	1.41	1.52	1.16	1.69	0.60	2.00	0.00						
11	B	10.0	11.3	2.74	3.92	1.52	2.78	0.91	2.09	0.71	1.97	0.55	1.73	0.37	1.63						
10	C	3.33	5.33	2.18	4.18	1.67	3.67	1.31	3.31	1.11	3.11	0.94	2.94	0.67	2.67						
	A	8.00	9.09	2.52	2.89	1.80	1.86	1.69	0.88	1.68	0.71	2.00	0.00			$s = 8$ 2.00	0.00				
9	B	8.00	9.09	2.22	3.18	1.29	2.38	0.80	1.77	0.65	1.74	0.52	1.48			0.46	1.54				
8	C	3.20	5.20	2.07	4.07	1.60	3.60	1.24	3.24	1.07	3.07	0.89	2.89			0.80	2.80				
	A	6.00	6.80	2.10	1.81	1.77	1.17	2.00	0.00	2.00	0.00										
7	B	6.00	6.80	1.67	2.24	1.07	1.87	0.71	1.29	0.60	1.40										
6	C	3.00	5.00	1.90	3.90	1.50	3.50	1.29	3.29	1.00	3.00										
	A	4.00	4.22	2.00	0.00	2.00	0.00														
5	B	4.00	4.22	1.20	0.80	0.89	1.11														
4	C	2.67	4.67	1.60	3.60	1.33	3.33														
	A	2.00	0.00																		
3	B	2.00	0.00																		
	C	2.00	4.00																		

*Values in bold face type are for the first value of n shown for the row; remaining values are for the second value of n.

$$V(\hat{g}_i - \hat{g}_j) = 2(a^0 - a^{|i-j|})[\sigma_s^2 + (\sigma_e^2/r)].$$

Expressions for a^0 and $a^{|i-j|}$ derived in the appendix at the end of this paper could be used to evaluate this variance for each pair of values of i and j. We shall consider only the average of this variance over the $n(n - 1)/2$ possible comparisons of general combining abilities. Using (3.4), this average variance is,

$$\text{Av. } V(\hat{g}_i - \hat{g}_j) = 2\left\{\frac{na^0}{n - 1} - \frac{1}{2s(n - 1)}\right\}[\sigma_s^2 + (\sigma_e^2/r)].$$

With $s/2$ common testers, the variance of all comparisons will be

$$V(\hat{g}_i - \hat{g}_j) = 4/s[\sigma_{s(t)}^2 + (\sigma_e^2/r)].$$

Again making the big assumptions that the general combining ability of each line with the common testers is the same quantity as the general combining ability of each line with all other inbred lines and that $\sigma_{s(t)}^2 = \sigma_s^2$, the average efficiency of the partial diallel cross relative to common testers in estimating general combining abilities is $E = 4(n - 1)/(2nsa^0 - 1)$. We have taken the total number of crosses as well as the total number of experimental plots the same for both designs. Values of E can be calculated from Table 5 and are shown in Table 7. When $s = 2$, $E = 4/(n + 1)$ and when $s = n - 1$ (the complete diallel cross), $E = 2(n - 2)/(n - 1)$. The partial diallel cross is very inefficient compared with common testers when $s = 2$ for all values of n above 3. The complete diallel cross is always at least as efficient as common testers and tends to be twice as efficient as n becomes large. The partial diallel cross with intermediate values of s is often at least as efficient as common testers. If having slightly unequal standard errors for the different comparisons is not a serious disadvantage, the partial or complete diallel cross is a better design for estimating general combining abilities than common testers providing that a sufficient number of crosses can be made so that s/n is a reasonably large fraction. Providing $s > 2$, the standard errors for comparisons can be estimated by using the analysis of variance shown in Table 1 for the partial (and complete) diallel cross and the usual analysis of variance for common testers.

For a fixed total number of crosses, the partial diallel cross does permit more lines to be included in the experiment than does the complete diallel cross. In selecting lines for general combining ability, the resulting increase in the selection intensity will often more than compensate for the decreased accuracy with which each line is assessed.

7. *Summary*

A diallel cross among n inbred lines with no parental or reciprocal crosses involves a total of $n(n - 1)/2$ crosses. Clearly, this number

TABLE 7

EFFICIENCY* OF PARTIAL DIALLEL CROSS RELATIVE TO COMMON TESTERS IN ESTIMATING GENERAL COMBINING ABILITIES

	s									
n	2	3	4	5	6	7	10	20	50	100
101, 100	**0.04**	0.11	**0.19**	0.29	**0.40**	0.51	**0.84**	**1.52**	**1.93**	**1.98**
75, 74	**0.05**	0.15	**0.25**	0.38	**0.51**	0.64	**1.00**	**1.51**	**1.95**	[s = 74, 1.97]
51, 50	**0.08**	0.21	**0.35**	0.52	**0.68**	0.85	**1.22**	**1.77**	**1.96**	
21, 20	**0.18**	0.46	**0.71**	1.00	**1.18**	1.38	**1.68**	**1.90**		
11, 10	**0.33**	0.80	**1.10**	1.44	**1.56**	1.69	**1.80**			
9, 8	**0.40**	0.93	**1.24**	1.55	**1.64**	1.71	[s = 8, 1.75]			
7, 6	**0.50**	1.14	**1.40**	1.60	**1.67**					
5, 4 3	**0.67** **1.00**	1.33	**1.50**							

* Values in boldface type are for the first value of n shown for the row; remaining values are for the second value of n.

increases rapidly with n. With limited facilities, this may mean that a complete diallel cross can only be made among a rather small number of inbred lines. The design presented in this paper allows a large number of inbred lines to be studied by performing only a sample of all the possible crosses among them. The efficiency of the design for the estimation of the variances of the general and specific combining abilities of the lines, for the prediction of the yielding capacities of the various crosses and for the estimation of the general combining ability of each of the lines, is discussed. The design is shown often to be more efficient than other designs that have been proposed.

8. *Acknowledgement*

We are indebted to Aaron Booker and Mary Clem for computations of a^0 values and other quantities.

APPENDIX

Inversion of Matrix A

A is a real symmetric circulant matrix and so can be inverted explicitly without much difficulty. Consider the general circulant matrix

$$\mathbf{A} = \begin{Bmatrix} a_0 & a_1 & \cdots & a_{n-1} \\ a_{n-1} & a_0 & \cdots & a_{n-2} \\ \vdots & \vdots & & \vdots \\ a_1 & a_2 & \cdots & a_0 \end{Bmatrix}.$$

Let $w_j = \exp. \{j(2\pi i/n)\}$, $(j = 1, 2, \cdots, n)$, be the n-th roots of unity. Then

$$\mathbf{A}\begin{bmatrix} 1 \\ w_j \\ w_j^2 \\ \vdots \\ w_j^{n-1} \end{bmatrix} = (a_0 + a_1 w_j + \cdots + a_{n-1} w_j^{n-1}) \begin{bmatrix} 1 \\ w_j \\ w_j^2 \\ \vdots \\ w_j^{n-1} \end{bmatrix}.$$

Therefore, the characteristic roots of **A** are

$$\lambda_j = a_0 + a_1 w_j + \cdots + a_{n-1} w_j^{n-1}, \qquad j = 1, 2, \cdots, n. \tag{1}$$

We can invert these equations to obtain the a's in terms of the λ's,

$$a_0 = \frac{1}{n}[\lambda_1 + \lambda_2 + \cdots + \lambda_n],$$

and

$$a_j = \frac{1}{n} [\lambda_1 w^j_{n-1} + \lambda_2 w^j_{n-2} + \cdots + \lambda_n], \qquad j = 1, 2, \cdots, n - 1.$$

Now, $\mathbf{A}^{-1}$ is also a circulant matrix and has characteristic roots $1/\lambda_j (j = 1, 2, \cdots, n)$. Therefore its elements can be written

$$a^0 = \frac{1}{n} \left[\frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \cdots + \frac{1}{\lambda_n} \right]$$

and

$$a^j = \frac{1}{n} \left[\frac{1}{\lambda_1} w^j_{n-1} + \frac{1}{\lambda_2} w^j_{n-2} + \cdots + \frac{1}{\lambda_n} \right], \quad j = 1, 2, \cdots, n - 1. \quad (2)$$

Since the matrix $\mathbf{A}$ is real and symmetric, its characteristic roots λ_j and the elements of the inverse matrix are both real. Substituting $a_0 = s$, $a_1 = a_2 = \cdots = a_{k-1} = a_{k+s} = a_{k+s+1} = \cdots = a_{n-1} = 0$ and $a_k = a_{k+1} = \cdots = a_{k+s-1} = 1$ in (1) and taking real parts,

$$\lambda_j = s - \frac{\sin (n - s) \dfrac{j\pi}{n}}{\sin \dfrac{j\pi}{n}} \qquad j = 1, \cdots, n - 1,$$

and (3)

$$\lambda_n = 2s.$$

Taking real parts on both sides of equations (2),

$$a^0 = \frac{1}{n} \left[\frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \cdots + \frac{1}{\lambda_n} \right],$$

and

$$a^j = \frac{1}{n} \sum_{l=1}^{n} \frac{1}{\lambda_l} \cos \frac{j(n - l)}{n} 2\pi, \qquad j = 1, 2, \cdots, n - 1. \qquad (4)$$

The elements of the inverse matrix can therefore be obtained by substituting the equations (3) for the λ's into the equations (4). This method was used to compute the values of a^0 in Table 5. For small values of n, the methods available for inverting general matrices may be preferable to the special method outlined above for inverting circulant matrices.

REFERENCES

1. Cockerham, C. C. [1954]. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics 39*, 859–82.
2. Comstock, R. E. and Robinson, H. F. [1948]. The components of genetic variance

in populations of biparental progenies and their use in estimating the average degree of dominance. *Biometrics 4*, 254–66.

3. Comstock, R. E. and Robinson, H. F. [1952]. Estimation of average dominance of genes. Chapter 30 of *Heterosis*, edited by J. W. Gowen. Iowa State College Press, Ames.
4. Finney, D. J. [1958a]. Statistical problems of plant selection. *Bull. de l'Inst. Internat. de Stat. 36* (3), 242–68.
5. Finney, D. J. [1958b]. Plant selection for yield improvement. *Euphytica 7*, 83 106.
6. Finney, D. J. [1960]. A simple example of the external economy of varietal selection. *Bull. de l'Inst. Internat. de Stat., 37* (3), 91–106.
7. Fisher, R. A. [1918]. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinb. 52*, 399–433.
8. Griffing, B. [1956a]. A generalized treatment of the use of diallel crosses in quantitative inheritance. *Heredity 10*, 31–50.
9. Griffing, B. [1956b]. Concept of general and specific combining ability in relation to diallel crossing systems. *Aust. J. Biol. Sci. 9*, 463–93.
10. Hayman, B. I. [1954a]. The analysis of variance of diallel tables. *Biometrics 10*, 235–44.
11. Hayman, B. I. [1954b]. The theory and analysis of diallel crosses, *Genetics 39*, 789-809.
12. Jensen, S. [1959]. *Combining ability of unselected inbred lines of corn from incomplete diallel and top-cross tests.* Unpublished Ph.D. Thesis, Iowa State University.
13. Jinks, J. L. and Hayman, B. I. [1953]. The analysis of diallel crosses *Maize Genetics Coop. News Letter 27*, 48–54.
14. Kempthorne, O. [1953]. A class of experimental designs using blocks of two plots. *Ann. Math. Stat. 24*, 76–84.
15. Kempthorne, O. [1956]. The theory of the diallel cross. *Genetics 41*, 451–59.
16. Kempthorne, O. [1957]. *An introduction to genetic statistics.* John Wiley and Sons, New York.
17. Sprague, G. F. and Tatum, L. A. [1942]. General vs. specific combining ability in single crosses of corn. *J. Amer. Soc. Agron. 34*, 923–32.
18. Zoellner, J. A. and Kempthorne, O. [1954]. Incomplete block designs with blocks of two plots. *Agricultural Experiment Station Research Bulletin 418*, Iowa State College, Ames, Iowa.

SENSORY TESTING BY TRIPLE COMPARISONS

G. T. Park

T. Hedley and Co. Ltd., Gosforth, Newcastle upon Tyne, England.

1. INTRODUCTION

In sensory test work, the number of samples a judge can assess at any one time is often limited by sensory fatigue and confusion of stimuli. This necessitates the use of incomplete block designs. Further, the difficulty of obtaining satisfactory quantitative measures of treatment effects usually entails assessment by ranking. Thus there is need for a rank analysis of incomplete block designs of a versatility equivalent to that available for variables satisfying the conditions of analysis of variance. In addition, any such analysis should provide the experimenter with a check as to whether he can regard the treatments concerned as lying on a linear scale of acceptability. For, with sensory assessment, cases can and do arise where the relations between calculated average scale positions disagree with individual direct comparisons.

These needs have been met in the case of blocks of size two, by the various paired comparison procedures developed in recent years [2, 4–6; 13, 14, 18, 17]. Although this is undoubtedly the most important case, instances arise where blocks of size three or more are a practical proposition and hold out the possibility of increased efficiency.

Unfortunately, little work seems to have been done in the general case of blocks of size k. The result published by Durbin [7] allows an overall significance test of treatment differences on balanced incomplete block ranking experiments, whilst the rankit method of Fisher [10] has been used by some workers, but no generally accepted method (of adequate versatility) has been available for blocks of size other than two. The method recently published by Pendergrass and Bradley [3, 15] which offers a comprehensive treatment of the case $k = 3$ is therefore of immediate interest.

This paper is to assist evaluation of the Bradley-Pendergrass method by contrasting results from triple comparison testing with results obtained using paired comparisons. Two sets of data are discussed. One set is from margarine cocktail stick taste tests, where sensory fatigue is known to be marked. The other set is from toilet soap sniff tests, where sensory fatigue is much less of a problem.

2. BRADLEY-PENDERGRASS TRIPLE COMPARISON ANALYSIS

The members of each of the $\binom{t}{3}$ triplets formed by the set of t treatments are ranked in order of acceptability by n judges.[1] Analysis proceeds as follows.

The Model

The model used is an extension of that successfully developed by Bradley and Terry for paired comparisons, [2, 4–6, 8, 11]. It is supposed that the treatments $T_1, \cdots, T_t$ are associated with parameters $\pi_1, \cdots, \pi_t$ which satisfy the following conditions:

$$\pi_i \geq 0, \qquad \sum_{i=1}^{t} \pi_i = 1,$$

$$P(T_i > T_j > T_k) = \pi_i^2 \pi_j / \Delta_{ijk},$$

$$\Delta_{ijk} = \pi_i^2(\pi_j + \pi_k) + \pi_j^2(\pi_i + \pi_k) + \pi_k^2(\pi_i + \pi_j),$$

$$i \neq j \neq k; \qquad i, j, k = 1, \cdots, t,$$

where $P(T_i > T_j > T_k)$ represents the probability that in the block containing treatments T_i, T_j, T_k, treatment i is rated top, treatment j central and treatment k bottom, in acceptability.

Maximum Likelihood Estimators of Parameters π_i

The maximum likelihood estimators $p_1, \cdots, p_t$ of the parameters $\pi_1, \cdots, \pi_t$ are given by the equations below:

$$a_i/p_i = n \sum_{\substack{j<k \\ j,k \neq i}} \{[2p_i(p_j + p_k) + p_j^2 + p_k^2]/D_{ijk}\},$$

$$\sum_i p_i = 1, \qquad i, j, k = 1, \cdots, t$$

where a_i = twice the total number of first places + the total number of second places given to T_i and

$$D_{ijk} = p_i^2(p_j + p_k) + p_j^2(p_i + p_k) + p_k^2(p_i + p_j).$$

In most cases these can be solved by one or two iterations using as initial estimates the values of p_i found from the quadratic equations:

$$p_i^2[a_i(2t^2 - 6t + 6) - n(2t^4 - 11t^3 + 22t^2 - 19t + 6)]$$
$$+ p_i[a_i(2t - 6) - n(t^3 - 4t^2 + 5t - 2)] + 2a_i = 0, \qquad i = 1, \cdots, t.$$

[1] Extension of the analysis to the general case of n_{ijk} rankings on the triplet $T_i, T_j, T_k, n_{ijk} \geq 0$, $i < j < k, i, j, k = 1 \cdots t$, is readily made and results quoted by the authors in [15].

Variance—Covariance Matrix of Estimators p_i

Expressions for the variances and covariances are presented by Bradley and Pendergrass for large n. These are particularly awkward to compute, however, and the authors provide an approximation which should suffice for practical purposes. If $\pi_1 \cdots \pi_t$ are near equality, then

$$\operatorname{Var}(\sqrt{n}p_i) = 3(t-1)/t^4(t-2), \qquad i = 1, \cdots, t,$$
$$\operatorname{Cov}(\sqrt{n}p_i, \sqrt{n}p_j) = -3/t^4(t-2), \quad i \neq j, \quad i, j = 1, \cdots, t.$$

Goodness of Fit

The model can be tested by comparing observed frequencies f_{ijk} with expected frequencies f'_{ijk}. If f_{ijk} = number of times treatments T_i, T_j, T_k were ranked in that order of acceptability, and

$$f'_{ijk} = np_i^2p_j/D_{ijk}, \qquad i \neq j \neq k, \qquad i, j, k = 1, \cdots, t,$$

then

$$\sum_{\substack{i \neq j \neq k \\ i,j,k}} \frac{(f_{ijk} - f'_{ijk})^2}{f'_{ijk}}$$

is distributed as chi-square with $\left[5\binom{t}{3} - \left(t - 1\right)\right]$ degrees of freedom for large n.

Scaling of Treatments

If the data fit the model, then the authors suggest that the treatments be regarded as lying at points $\ln p_i$ on a linear scale of acceptability. This is by analogy with the paired comparison model of Bradley and Terry. In that case, the parameters π_i can be shown to be related to a metric on which the mean response for treatment T_i is $\ln \pi_i$.

Overall Tests of Main Effects and Interaction

An overall test for main effects is available if required. For large n,

$$Z = 2n\binom{t}{3} \ln 6 + 2 \sum_i a_i \ln p_i - 2n \sum_{i<j<k} \ln D_{ijk}$$

is distributed as chi-square with $(t - 1)$ degrees of freedom. Further, if the experiment has been repeated over a number of different groups of judges or in different circumstances, interaction can be tested using the result that: $Z^c - Z$ is distributed as $\chi^2(g - 1)(t - 1)$ where $Z^c = \sum_{u=1}^{g} Z_u$, g = no. of groups, $Z_u = Z$ above, computed for uth group, $u = 1 \cdots g$.

Relative Efficiency of Paired Comparisons and Triple Comparisons

Under the assumption that differences observed by judges between the three component pairs of each triplet are the same as those which would be observed in the equivalent paired comparisons, then the efficiency of triple comparisons to paired comparisons is 9/4 in terms of incomplete block rankings (or $\frac{3}{2}$ in terms of treatment replications, as three assessments are made per triplet ranking against two per pair ranking).

3. PAIRED COMPARISON ANALYSES

As the various paired comparison procedures give closely similar results [1, 12], any one of them could be used as the basis for assessment of the triple comparison technique. However, that of Bradley and Terry is an obvious choice because of its relationship to the triple comparison method. In addition, the results from Scheffé's method [17] are given, as results from this method had to be obtained for routine reporting of the paired comparison legs. It must be expected, of course, that the results of these two analyses will agree more closely with each other than with the triple comparison results, as they are derived from the same basic data, whereas the triple comparison data is distinct and independent.

TABLE I

RESULTS OF TEST I

Statistic	Paired Comparisons		Triple Comparisons
	Scheffé	B-Terry	B-Pendergrass
Response Ratings[2]	α_i	$\ln p_i$	$\ln p_i$
Product A	-0.55	-0.42	-0.06
Product B	0.22	0.28	-0.34
Product C	0.36	0.43	0.40
Product D	0.45	0.28	0.50
Product E	-0.48	-0.57	-0.50
S.E. of Diff.	± 0.48	± 0.52	± 0.33
Goodness of Fit	$\chi^2(6) = 2.3$	$\chi^2(6) = 5.7$	$\chi^2(46) = 47.5$
Treatment Diffs.	$\chi^2(4) = 8.2$	$\chi^2(4) = 5.7$	$\chi^2(4) = 13.8$

4. MARGARINE COCKTAIL STICK TASTE TESTS

It was found that guidance on the likely results of full-scale consumer research studies could be obtained by taste testing small cylinders of margarine from cocktail sticks. Two such tests were done for routine purposes by paired comparisons and for experimental purposes by triple comparisons.

Test I. In Test I, judges examined five margarines A–E. Each of 60 judges made one paired comparison to give 3 rankings per ordered pair. Each of a further 120 judges made one triple comparison to give 2 rankings per ordered triplet.

Test II. In Test II, judges examined three margarines L–N. Each of 138 judges made one paired comparison to give 23 rankings per ordered pair. Each of a further 126 judges made one triple comparison to give 21 rankings per ordered triplet.

Paired comparison and triple comparison ratings are seen to be in close agreement on Test II. On Test I agreement is not so good, there being an indication that the methods give different measures of the relative acceptability of Products *A* and *B*. The goodness of fit chi-squares show that both of the paired comparison models and the triple comparison model gave acceptable fits to their respective data.

On Test II, the relative sensitivity of the three procedures on a per ranking basis is indicated by the treatment chi-squares as they stand, for the number of pair rankings approximated the number of

TABLE 2

RESULTS OF TEST II

Statistic	Paired Comparisons		Triple Comparisons
	Scheffé	B-Terry	B-Pendergrass
Response Ratings[2]	α_i	$\ln p_i$	$\ln p_i$
Product L	−0.45	−0.38	−0.44
Product M	0.55	0.62	0.56
Product N	−0.10	−0.24	−0.12
S.E. Of Diff.	±0.46	±0.58	±0.41
Goodness of Fit	$\chi^2(1) = 0.0$	$\chi^2(1) = 0.2$	$\chi^2(3) = 1.3$
Treatment Diffs.	$\chi^2(2) = 4.8$	$\chi^2(2) = 3.2$	$\chi^2(2) = 6.0$

[2] All sets of response ratings have been adjusted to total zero and range one.

triplet rankings. In the case of Test I, it is necessary to double the paired comparison chi-squares before comparing them with the triple comparison chi-square, as the number of pair rankings was only one half the number of triplet rankings. In each case, there is no indication that triple comparisons were more efficient than paired comparisons. It is of interest that application of Durbin's chi-square test of treatment differences to the triple comparison data from Tests I and II gave values close to those found under the Bradley-Pendergrass analysis ($\chi^2(4) = 13.4$, $\chi^2(2) = 6.0$).

5. TOILET SOAP SNIFF TESTS

Two toilet soap sniff tests were done for routine purposes by paired comparisons and for experimental purposes by triple comparisons. *Test III* examined four toilet soap perfumes A–D. Each of 132 judges made one paired comparison to give 11 rankings per ordered pair.

Each of a further 131 judges made one triple comparison. Although the design arranged for rankings to be divided evenly over the 24 ordered triplets, an error in arranging record sheets resulted in the number of rankings varying as follows: (Order of presentation within triplet remained in approximate balance). Analysis of this test was therefore carried out using the extension of the method to the general case when the number of rankings differs between triplets.

	Triplet			
	ABC	ABD	ACD	BCD
No. of rankings	51	17	11	52

Test IV examined six toilet soap perfumes L–Q. Each of 180 judges made two successive paired comparisons (involving four different products). As there was no evidence that results from pairs sniffed first differed from those for pairs sniffed second, the results for all pairs were combined to give a total of 360 rankings, i.e. 12 rankings per ordered pair. Each of a further 120 judges made one triple comparison to give one ranking per ordered triplet.

On both tests, paired comparison and triple comparison ratings are in agreement within the limits of error.

On Test III, the paired comparison models fitted the paired comparison data, but the triple comparison model was not a satisfactory fit to the triple comparison data (goodness of fit chi-square significant

TABLE 3

Results of Test III

Statistic	Paired Comparisons		Triple Comparisons
	Scheffé	B-Terry	B-Pendergrass
Response Ratings	α_i	$\ln p_i$	$\ln p_i$
Product A	0.30	0.23	0.30
Product B	0.44	0.46	0.55
Product C	−0.56	−0.54	−0.40
Product D	−0.18	−0.15	−0.45
S.E. of Diff.[3]	±0.42	±0.50	±0.31
Goodness of Fit	$\chi^2(3) = 0.3$	$\chi^2(3) = 0.2$	$\chi^2(17) = 40.1$
Treatment Diffs.	$\chi^2(3) = 7.2$	$\chi^2(3) = 4.5$	$\chi^2(3) = 16.5$

[3] An average value is given for the S.E. of difference of triple comparison ratings as this figure varies between pairs of products due to differing numbers of rankings per triplet.

TABLE 4

Results of Test IV

Statistic	Paired Comparisons		Triple Comparisons
	Scheffé	B-Terry	B-Pendergrass
Response Ratings	α_i	$\ln p_i$	$\ln p_i$
Product L	0.30	0.37	−0.04
Product M	−0.55	−0.45	−0.36
Product N	0.45	0.30	0.16
Product O	−0.55	−0.63	−0.55
Product P	0.13	0.19	0.34
Product Q	0.22	0.22	0.45
S.E. of Diff.	±0.38	±0.29	±0.21
Goodness of Fit	$\chi^2(10) = 5.7$	$\chi^2(10) = 6.5$	$\chi^2(95) = 76.7$
Treatment Diffs.	$\chi^2(5) = 13.5$	$\chi^2(5) = 18.5$	$\chi^2(5) = 32.8$

at 1 percent level). On Test IV, all three models fitted their respective data.

Using the treatment difference chi-squares as before to compare sensitivity on a per ranking basis, it appears that, on both tests, considerably greater sensitivity was achieved with triple comparisons than with paired comparisons. (Note that on Test IV the number of triplet rankings was only one-third the number of pair rankings so that the triple comparison chi-square requires multiplication by three before contrasting with the paired comparison chi-squares). Application of Durbin's test of treatment differences to the triple comparison data from the balanced test, Test IV, gave $\chi^2(5) = 31.5$ in good agreement with that found under the Bradley-Pendergrass analysis.

6. ANALYSIS OF VARIANCE-TEST III

As the extension of the triple comparison analysis to the case of varying numbers per triplet was not available at the time, the triple comparison results from Test III were analysed by least squares [using dummy variables [16]]. As shown below, the results agreed very closely with those subsequently found by the triple comparison method.

	Products				S.E. of	Treatment
	A	B	C	D	Diff.	$\chi^2(3)$
Triple Comparison	0.30	0.55	−0.40	−0.45	±0.31	16.5
Least Squares	0.32	0.54	−0.40	−0.46	±0.31	17.1

7. DISCUSSION AND SUMMARY

These results suggest that the model on which the triple comparison analysis is based will prove appropriate to at least a proportion of sensory tests. Further, that in those cases where sensory fatigue is not marked, it may afford a valuable gain in efficiency over paired comparisons. It is pointed out, however, that each of the experiments described above compared products which were basically similar and which differed essentially in only one dimension.

In the author's experience, paired comparison tests involving treatment differences of this nature rarely give results inconsistent with the various paired comparison models. Lack of fit tends to occur when treatments differ importantly in more than one dimension or when marked judge × treatment interactions are likely to be present. As the triple comparison model is likely to be more vulnerable than the paired comparison models to the complications arising on these latter types of test, it may prove of limited applicability where treatments differ in kind rather than in degree.

ACKNOWLEDGMENT

The author is indebted to R. A. Bradley for commenting on the first draft of this paper and to Miss Anne Thirlwall for assistance with the computation.

REFERENCES

[1] Bliss, C. I., Greenwood, Mary L., and White, Edna S. [1956]. A rankit analysis of paired comparisons for measuring the effect of sprays on flavour. *Biometrics 12*, 381–403.

[2] Bradley, R. A. and Terry, M. E. [1952]. Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika 39*, 324–45.

[3] Bradley, R. A. and Terry, M. E. [1952]. The rank analysis of incomplete block designs II. The method for blocks of size three. *Appendix A, Bi-annual Report No. 4, Va. Agr. Exp. Sta.*

[4] Bradley, R. A. [1954]. Rank analysis of incomplete block designs II. Additional tables for the method of paired comparisons. *Biometrika 41*, 502–37.

[5] Bradley, R. A. [1954]. Incomplete block rank analysis. On the appropriateness of the model for a method of paired comparisons. *Biometrics 10*, 375–90.

[6] Bradley, R. A. [1955]. Rank analysis of incomplete block designs III. Some large-sample results on estimation and power for a method of paired comparisons. *Biometrika 42*, 450–70.

[7] Durbin, J. [1951]. Incomplete blocks in ranking experiments. *Brit. Jour. Psych. 4*, 85–90.

[8] Dykstra, O. [1956]. A note on the rank analysis of incomplete block designs—Applications beyond the scope of existing tables. *Biometrics 12*, 301–06.

[9] Dykstra, O. [1958]. Factorial experimentation in Scheffé's analysis of variance for paired comparisons. *J.A. S.A.* 53, 529–42.

[10] Fisher, R. A. and Yates, F. [1953]. *Statistical Tables for Biological, Agricultural and Medical Research.* Fourth Edition, Oliver & Boyd, Edinburgh, 25, 76.

[11] Hopkins, J. W. [1954]. Incomplete block rank analysis: some taste test results. *Biometrics 10*, 391–99.

[12] Jackson, J. E. and Fleckenstein, Mary [1957]. An evaluation of some statistical techniques used in the analysis of paired comparison data. *Biometrics 13*, 51–64.

[13] Mosteller, F. [1951 a]. Remarks on the method of paired comparisons I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika 16*, 3–9.

[14] Mosteller, F. [1951 b]. Remarks on the method of paired comparisons II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika 16*, 203–18.

[15] Pendergrass, R. N. and Bradley, R. A. [1960]. Ranking in triple comparisons. *Contributions to Probability and Statistics,* edited by I Olkin et. al., Stanford University Press, Stanford, California.

[16] Quenouille, M. H. [1950]. *Introductory Statistics,* 148–54.

[17] Scheffé, H. [1952]. An analysis of variance for paired comparisons. *J.A.S.A.* *47,* 381–400.

[18] Thurstone, L. L. [1927]. Psychophysical analysis. *Amer. J. Psychol 38,* 368.

RAPID CHI-SQUARE TEST OF SIGNIFICANCE FOR THREE-PART RATIOS[1]

Charles E. Gates and Benjamin H. Beard[2]
University of Minnesota, St. Paul, Minnesota, U.S.A.

INTRODUCTION

In the study of qualitative characters, the number of individuals distributed among mutually exclusive categories is commonly compared with a distribution predicated on genetic theory. If the observed numbers in the various classes deviate from the theoretical, it is desirable to be able to reject the hypothetical distribution at a determinable probability level. A statistic normally used in ascertaining this goodness-of-fit is chi-square. Frequently, the number of such comparisons may be large and the computations necessary for each segregating family become laborious. In these instances an accurate, fast technique would prove useful. This paper presents a rapid, graphical procedure for testing goodness-of-fit to ratios with three classes. These graphical chi-square tests of significance are particularly appropriate for testing lack of fit to three-part ratios for those organisms where it is feasible to take random samples of some constant number of individuals from segregating progenies. However, by constructing a series of graphs, it is practical to use this method with variable sample size if the number of uses is sufficiently large. Graphs to be used with the 1 : 2 : 1 ratio have been drawn for a variety of values of n and have been used successfully in the field for determining lack of fit.

Although the justification for, as well as the illustrated examples, have been drawn from genetics, uses in other fields can be visualized. One particular application outside the field of genetics could be in evaluating the results of a sample survey where a response can be categorized into three mutually exclusive classes and one is interested in ascertaining the goodness-of-fit to a 1 : 1 : 1 ratio.

ILLUSTRATION

Suppose the numbers of individuals falling in three mutually exclusive classes are x_1, x_2 and x_3, where

$$x_1 + x_2 + x_3 = n. \tag{1}$$

[1] Paper No. 4346, Scientific Journal Series, Minnesota Agricultural Experiment Station.

[2] Crops Research Division, Agricultural Research Service, U. S. Department of Agriculture.

The critical region for a particular level of significance can be defined by an ellipse drawn on rectangular coordinate paper with x_2 and x_1 (or x_3) the ordinate and the abscissa, respectively. Figure 1 was prepared for the 1 : 2 : 1 ratio, the three ellipses indicating demarcation between the .10, .05 and .01 levels of significance.

To illustrate the use of this graph, suppose the segregates from individual F_1 plants from a cross of two barley varieties differing in number of rows of kernels on the spike are classified and counted. In a cross of this kind (two-row VV × six-row vv) the heterozygote (Vv) is phenotypically distinguishable from either homozygote and, if conditioned by one gene, the F_2 should segregate in a 1 : 2 : 1 ratio. Consider the following examples wherein x_1, x_2 and x_3 refer respectively to the numbers in the VV, Vv and vv classes:

Example	x_1	x_2	x_3	n
1	20	16	14	50
2	16	22	12	50
3	7	33	10	50
4	15	33	2	50

Locating the points of intersection of x_1 and x_2 (or x_3 and x_2), as measured along the abscissa and ordinate, respectively, the probabilities from the χ^2 statistic can be immediately ascertained as $.05 > P > .01$, $P > .10$, $.10 > P > .05$ and $P < .01$ for examples 1, 2, 3 and 4, respectively.

THEORETICAL DEVELOPMENT FOR AN ARBITRARY THREE-PART RATIO

Genetic theory or other considerations predicate that three classes corresponding to those indicated by (1) should have the proportions π_1, π_2 and π_3, respectively. It can readily be seen that

$$E(x_i) = n\pi_i\ , \quad \text{Var}\,(x_i) = n\pi_i(1 - \pi_i), \quad \text{Cov}\,(x_i\ , x_j) = -n\pi_i\pi_j\ . \qquad (2)$$

It is well known (see, e.g., Cochran [1952]) that

$$[x_1 - n\pi_1\ x_2 - n\pi_2]\begin{bmatrix} n\pi_1(1 - \pi_1) & -n\pi_1\pi_2 \\ -n\pi_1\pi_2 & n\pi_2(1 - \pi_2)\end{bmatrix}^{-1}\begin{bmatrix} x_1 - n\pi_1 \\ x_2 - n\pi_2\end{bmatrix} \qquad (3)$$

will be distributed asymptotically as chi-square with 2 degrees of freedom. Expression (3) may be re-written as

$$[(x_1 - n\pi_1)^2\pi_2(1 - \pi_2) + 2\pi_1\pi_2(x_1 - n\pi_1)(x_2 - n\pi_2) \\ + (x_2 - n\pi_2)^2\pi_1(1 - \pi_1)]/(n\pi_1\pi_2\pi_3). \qquad (4)$$

Now (4) is algebraically identical to the chi-square goodness-of-fit statistic

$$\chi^2_{2,\alpha} = \sum_{i=1}^{3} [(x_i - n\pi_i)^2]/n\pi_i = \left[\sum_{i=1}^{3} x_i^2/n\pi_i\right] - n \tag{5}$$

indicating, incidentally, why the latter has its well known distribution.

Equation (5) can therefore be seen to represent an ellipse, not in standard form, in the x_1, x_2 plane, with the center at

$$(n\pi_1\ , n\pi_2). \tag{6}$$

The acceptance region of the usual goodness-of-fit test is the region inside this ellipse.

In order to plot ellipse (5), it is necessary to determine the slopes of the minor and major axis as well as their lengths. If the x_1, x_2 axes are rotated so that ellipse (5) will be in standard form, the slope of the rotated x_1 axis is

$$m = R + \sqrt{R^2 + 1} \tag{7}$$

where

$$R = (\pi_1 - \pi_2)\pi_3/2\pi_1\pi_2\ .$$

Rotating the ellipse, (5) can be written

$$[A(x_1 - n\pi_1)^2/\chi^2_{2,\alpha}] + [C(x_2 - n\pi_2)^2/\chi^2_{2,\alpha}] = 1 \tag{8}$$

where

$$A = [\pi_2(1 - \pi_2) + 2m\pi_1\pi_2 + \pi_1(1 - \pi_1)m^2]/n\pi_1\pi_2\pi_3(1 + m^2), \tag{9}$$

$$C = [\pi_2(1 - \pi_2)m^2 - 2m\pi_1\pi_2 + \pi_1(1 - \pi_1)]/n\pi_1\pi_2\pi_3(1 + m^2). \tag{10}$$

The semi-diameters of the minor and major axes are therefore

$$\chi_{2,\alpha}/\sqrt{A}, \qquad \chi_{2,\alpha}/\sqrt{C}, \tag{11}$$

respectively, where A, $C > 0$.

The slope of the rotated axis, as indicated by (7), is independent of sample size, n, as well as significance level. On the other hand, the location of the center of an ellipse, as shown by (6), is a function of n, but is independent of significance level. The lengths of the major and minor axes, given by (11), are a function of both sample size and significance level. Hence a family of ellipses with a common center and slope can be constructed for given n by varying significance level. In using these ellipses, it is recommended that the expected number in any class be the customary minimum of 5 or 10 because the distribution of (3) is only asymptotically distributed as χ^2.

CONSTRUCTION OF ELLIPSES FOR THE 1:2:1 RATIO

Substitution of 1/4, 1/2 and 1/4 for π_1, π_2 and π_3, respectively, yields the following basic quantities:

$$R = -1/4, \quad m = 0.780776, \quad 1 + m^2 = 1.609612,$$

$$A = 11.123103/n, \qquad C = 2.876897/n.$$

Making use of (6), the center of the 1 : 2 : 1 ellipse, is located at

$$(n/4, n/2). \tag{12}$$

Substituting the basic quantities in (11), one-half the lengths of the minor and major axes are found to be

$$.2998\sqrt{n}\chi_{2,\alpha} \quad \text{and} \quad .5896\sqrt{n}\chi_{2,\alpha}, \tag{13}$$

respectively. An ellipse can be readily constructed for each n in the range of interest using a Dietzgen ellipsograph or various simple, but accurate construction techniques such as the concentric-circle method for major and minor diameters. French and Vierck [1958], a recent reference, demonstrate various methods of ellipse construction.

To illustrate the procedure more specifically, consider ellipses for $n = 50$, where the critical regions are defined by the 0.10, 0.05 and 0.01 levels of significance. Making use of (12) and (13), one obtains These ellipses were plotted in Figure 1.

Significance level	Slope of rotated x_1 axis	Location of center		Semi-diameter length	
		x_1	x_2	x_1	x_2
.10	.781	12.50	25.00	4.55	8.94
.05	.781	12.50	25.00	5.20	10.22
.01	.781	12.50	25.00	6.43	12.65

CONSTRUCTION OF ELLIPSES FOR OTHER THREE-PART RATIOS

Calculation of ellipses for defining critical regions for other three-part ratios can be carried out in analogous fashion to the ellipses for the 1 : 2 : 1 ratio. To eliminate the necessity of deriving intermediate calculations, the slopes, center, and lengths of major and minor diameters for some of the more common three-part genetic ratios and the 1 : 1 : 1 ratio are given in Table 1. These quantities are expressed in terms of arbitrary n and significance level. To obtain figures for graph-

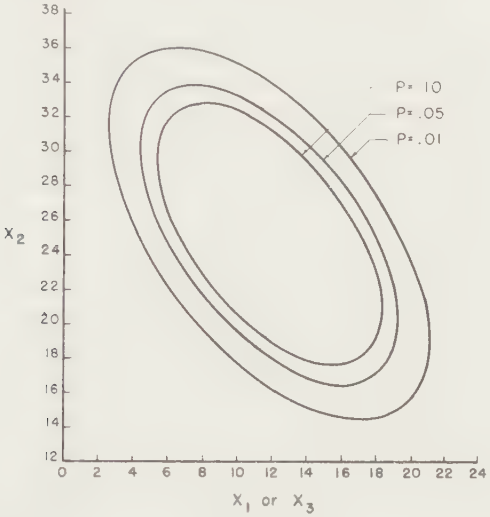


FIGURE 1

χ^2 TEST FOR A 1:2:1 RATIO

x_1, x_2 and x_3 are the observed numbers of progenies in the *VV*, *Vv* and *vv* classes, respectively. Intersection of x_1 and x_2 (or x_3 and x_2) outside an ellipse indicates significance at the P = .01, .05 and .10 levels of significance for sample size n = 50.

TABLE 1

SLOPES, CENTERS AND SEMI-DIAMETERS OF ELLIPSES DEFINING CRITICAL REGIONS FOR VARIOUS THREE-PART RATIOS.

Genetic Ratio*	Slope of minor axis of ellipse relative to x_1 axis	Center		Length of semi-diameters	
		x_1	x_2	x_1	x_2
3:9:4	0.650	$3n/16$	$9n/16$	0.289θ‡	0.561θ
6:9:1	0.973	$3n/8$	$9n/16$	0.171θ	0.672θ
3:9:3	0.721	$n/5$	$3n/5$	0.271θ	0.571θ
3:9:1	0.895	$3n/13$	$9n/13$	0.186θ	0.597θ
3:12:1	0.883	$3n/16$	$3n/4$	0.168θ	0.558θ
1:14:1	0.638	$n/16$	$7n/8$	0.154θ	0.380θ
1:2:1	0.781	$n/4$	$n/2$	0.300θ	0.590θ
1:1:1†	1.000	$n/3$	$n/3$	0.408θ	0.408θ

*π_2 was written as the largest component of each ratio. This is somewhat at variance from usual genetic terminology.

†non-genetic

‡$\theta = \sqrt{n}\, \chi_{2,\alpha}$

ing, it is necessary only to substitute n and $\chi^2_{2,\alpha}$ and to plot the ellipses as previously described. For three-part ratios not shown, it is necessary to substitute appropriate values for π_1 , π_2 , and π_3 in (6), (7), (9) and (10).

It would be possible to rewrite (4) in terms of the proportion $p_i = x_i/n$, resulting in corresponding changes in the above results. For a given genetic ratio and significance level, a family of ellipses could then be constructed with common center, the larger samples leading to the smaller ellipses. Such graphs would prove in some respects more convenient to use than the procedure outlined above. However, this convenience would in some instances, be offset by the requirement of the calculation of two of the p_i in order to determine the level of significance. The decision as to the better procedure depends on whether it is more desirable to have a minimum of separate graphs, but require supplementary computations, or to have more graphs with no computation necessary. Separate ellipses in either instance need not be constructed for each n as ellipses for n of similar size will be similar.

We are indebted to a referee for pointing out a more elegant approach to the derivation of the characteristics of the chi-square ellipses and to Donald Richter for helpful comments.

REFERENCES

Cochran, W. G. [1952]. The χ^2 test of goodness-of-fit. *Ann. Math. Stat. 23*, 315–45.
French, T. E. and Vierck, C. J. [1958]. *Graphic Science.* McGraw Hill, New York.

GENERATING UNBIASED RATIO AND REGRESSION ESTIMATORS

W. H. WILLIAMS

Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey, U. S. A.

1. INTRODUCTION

Information collected on a concomitant variate is often used in finite sampling theory to create more precise estimators of population characteristics. This supplementary information is obtained in addition to the characteristic under study and some aspects of it may be derived from sources other than the sample itself. It may be either qualitative or quantitative. For example, suppose that the variate under consideration in a sample survey is the number of dairy cattle per farm y and that at the time of the survey the number of grazing acres per farm x is also obtained. It may then be known from census data that the total number of grazing acres in the entire area is $N\mu_X$ and the mean per farm is μ_X. Analytically, we have a random sample of n pairs (y_i, x_i), $i = 1, \cdots, n$, from a population of size N and the population x-mean is known exactly. The problem is to estimate the population mean μ_Y.

A general class of estimators designed to utilize this supplementary information includes ratio and regression estimators. These estimators are described in textbooks on the subject, see for example Cochran [1953]. Additional developments have been presented by Hartley and Ross [1954], Nieto [1958] and Robson [1957].

The two classical ratio estimators are the ratio of means estimator $\tilde{y} = (\bar{y}/\bar{x})\mu_X$ and the mean of ratios estimator $\hat{y} = \mu_X \sum_{i=1}^{n} r_i/n$ where $\bar{y}$ and $\bar{x}$ are sample means and $r_i = y_i/x_i$. It is well known that these estimators are biased. The usual regression estimator is obtained by evaluating the least squares line of best fit $y = \bar{y} + b(x - \bar{x})$ at the point μ_X giving $\hat{y}_b = \bar{y} + b(\mu_X - \bar{x})$ as a regression estimator of μ_Y. This estimator is biased if the assumption of a linear model is not valid.

The generation of some exactly unbiased ratio and regression estimators is discussed in this paper. Specifically, we classify an estimator as of the regression type if it is invariant under location and scale changes in x and if it undergoes the same location and scale changes as the y variate. A ratio estimator has analogous properties but for scale changes only.

2. DERIVATION OF UNBIASED ESTIMATORS FOR SIMPLE RANDOM SAMPLING

To generate unbiased estimators, consider the following sampling procedure. At step one, select with equal probability one of all possible splits of the population into s mutually exclusive groups of size[1] n/k, i.e., $N = sn/k$. At the second stage, select randomly without replacement k of the groups from the total number of groups s of that particular split of the population. This gives a sample of size n.

Now consider the conditional distribution for a particular set of s groups. Attached to each of these groups there are characteristics[2] $\bar{y}^{(i)}$, $\bar{x}^{(i)}$, $b^{(i)}$, $i = 1, \cdots, s$, where $\bar{y}^{(i)}$ and $\bar{x}^{(i)}$ are means of the n/k units in the group and $b^{(i)}$ is as yet an unspecified function of the y and x of that group. For a given split and a random selection of groups, the expectations of $\bar{y}^i$ and $\bar{x}^i$, $i = 1, \cdots, k$, are μ_Y and μ_X respectively; that is, they are conditionally unbiased. Furthermore,

$$\left(1 - \frac{k}{s}\right)\frac{1}{k(k-1)}\sum_{i=1}^{k}(b^i - \bar{b})(\bar{x}^i - \bar{x}) \qquad (1)$$

is an unbiased estimator of Cov $(\bar{b}, \bar{x})$ where $\bar{b} = \sum_{i=1}^{k} b^i/k$.

Hence if $g = \bar{y} + \bar{b}(\mu_X - \bar{x})$ then $E(g) = \mu_Y - \text{Cov}\,(\bar{b}, \bar{x})$ and

$$T_k = \bar{y} + \bar{b}(\mu_X - \bar{x}) + \left(1 - \frac{n}{N}\right)\frac{1}{k(k-1)}\sum_{i=1}^{k}(b^i - \bar{b})(\bar{x}^i - \bar{x}) \qquad (2)$$

is a conditionally unbiased estimator of μ_Y. It is then unbiased unconditionally.

This approach is valid for any defined form of the coefficient $b^{(i)}$; T_k will remain unbiased. If $b^{(i)}$ has a form which is invariant under linear x and y transformation (say least squares form) then T_k is classified as a regression estimator. If $b^{(i)} = \bar{y}^{(i)}/\bar{x}^{(i)}$ (say), then T_k falls into the class of a ratio estimator.

This procedure is used to generate the unbiased estimators; in practice a simple random sample would be drawn and to compute T_k it would be split randomly into groups, see Section 7 for an example. The latter operation is equivalent to the generating procedure which allows a particular split-sample to arise in $\binom{s}{k}(N - n)!/[(n/k)!]^{s-k}$

[1] It is assumed that this relationship is true in terms of integers.

[2] Superscripts will be used to specify the groups. They will be used with parentheses when the reference is to the entire population of s groups and without parentheses when referring to the sample of k groups.

ways while splitting the simple random sample allows a particular split-sample to arise in only one way. The unbiasedness is preserved by either procedure.

The argument is easily generalized to p auxiliary variates.

3. SPECIFIC ILLUSTRATIONS IN SIMPLE RANDOM SAMPLING

A form of interest is

$$b^{(i)} = \frac{\sum_{j=1}^{n/k} (y_j - \bar{y}^{(i)})(x_j - \bar{x}^{(i)})}{\sum_{j=1}^{n/k} (x_j - \bar{x}^{(i)})^2}, \qquad i = 1, \cdots, s, \tag{3}$$

the least squares slope form.

In this case T_k bears much resemblance to $\hat{y}_b$ and might be thought of as possessing an additional component which is required to compensate for possible bias in $\hat{y}_b$. This is not exactly true of course, because the first two terms of T_k are not exactly the two terms of $\hat{y}_b$. However, in this case (3), it seems natural to make some remarks on the efficiency of T_k.

The variance of T_k depends very much on the form of the $b^{(i)}$ coefficients. In fact, until the form of $b^{(i)}$ is specified little can be said about the variance of T_k. One can imagine choices which would lead to poor efficiency indeed. However, T_k in this case has coefficients in the least squares slope form and it is natural to ask how it compares with $\hat{y}_b$ when a linear model is assumed, for then $\hat{y}_b$ has optimum variance properties. But with this assumption, $\hat{y}_b$ also possesses unbiasedness and the advantage of T_k is unbiasedness in situations in which $\hat{y}_b$ is not unbiased. However, one would like the efficiency of T_k to compare favorably even in this linear model case. So by assuming a linear model and a normal x-distribution, it is easily found that $V(\hat{y}_b)/V(T_k) = (n - 2)(n - 6)/(n - 3)(n - 4)$, $k = 2$ and $n > 6$. This expression is less than one but approaches one as n gets larger and, for example, when $n = 15, 25$ is equal to 0.89 and 0.95. Thus we see that one does not lose all the efficiency brought about by the use of an auxiliary variate and that $[V(T_k) - V(\hat{y}_b)]/V(\hat{y}_b)$ is $O(n^{-1})$.

Furthermore, the role of k also depends upon the choice of the $b^{(i)}$. For example, in the special case of the previous paragraph, if the number of groups is regarded as variable, $V(\hat{y}_b)/V(T_k)$ will be found to have a maximum at $k = \sqrt{n/3}$. Thus for this form of the $b^{(i)}$, the optimum number of groups is $\sqrt{n/3}$. Other forms of the $b^{(i)}$ would yield other results.

Another possible choice is

$$b^{(i)} = \sum_{j=1}^{n/k} y_j x_j \Big/ \sum_{j=1}^{n/k} x_j^2 \, .$$

In this form T_k is a ratio estimator and it is unbiased even if the linear relationship of y and x does not pass through the origin. But characteristically the variance will be inflated by such a relationship.

Next, if $b^{(i)} = \bar{y}^{(i)}/\bar{x}^{(i)} = r^{(i)}$, T_k will reduce to the form

$$T_k = \bar{r}\mu_X + \frac{Nk - n}{N(k - 1)} (\bar{y} - \bar{r}\bar{x}) \tag{4}$$

where $\bar{b}$ is denoted $\bar{r}$. It will be noted that when $k = n$, $T_k = y'$, the unbiased ratio estimator presented by Hartley and Ross [1954]. The efficiency of this form of T_k has been examined in detail by Goodman and Hartley [1958] and Robson [1957]. Robson presents an exact variance formula for finite populations.

Finally, consider $b^{(i)} = r^{(i)} = (k/n) \sum_{j=1}^{n/k} r_j$, $r_j = y_j/x_j$, then $\bar{b} = \bar{r} = \sum_{j=1}^{n} r_j/n$ which does not depend upon the particular split of the population. Now if, after substitution of this form into T_k , the estimator is averaged over all possible splits of the sample into groups of size n/k it will be found that the result is again the Hartley-Ross unbiased ratio estimator. This averaging process is indicated by a star, i.e., T_k^* .

Other forms could, of course, be considered.

4. STRATIFIED SAMPLING

Since a bias may be magnified relative to the standard deviation, stratified sampling may perhaps be regarded as the most important application of unbiased estimators. Their separate use within strata requires exact knowledge of the population strata means but is straightforward. We now develop a combined stratified estimator.

Consider L strata of size N_t , $t = 1, \cdots , L$ with $\sum_{t=1}^{L} N_t = N$, and again consider the sampling in two stages. At the first stage select with equal probability one of the possible splits of each stratum into s groups of size n_t/k, $t = 1, \cdots , L$. Then $N_t = sn_t/k$. At the second stage select k groups with equal probability and without replacement from each of the strata, giving a sample of size n_t in the t-th stratum, $\sum_{t=1}^{L} n_t = n$.

For a given split and a random selection of groups

$$\bar{y}_{st}^{i} = \sum_{t=1}^{L} (N_t/N)\bar{y}_t^{i} \quad \text{and} \quad \bar{x}_{st}^{i} = \sum_{t=1}^{L} (N_t/N)\bar{x}_t^{i}$$

are unbiased estimators of μ_Y and μ_X respectively, where $\bar{y}_t^i$ and $\bar{x}_t^i$ denote means of the i-th group in the t-th stratum. Also we can consider a coefficient $b_{st}^{(i)}$ which is as yet unspecified in form but utilizes the set of elements in the i-th group of all strata. For example,

$$b_{st}^{(i)} = \frac{\sum_{t=1}^{L}\sum_{j=1}^{n_t/k}(y_{tj} - \bar{y}_t^{(i)})(x_{tj} - \bar{x}_t^{(i)})}{\sum_{t=1}^{L}\sum_{j=1}^{n_t/k}(x_{tj} - \bar{x}_t^{(i)})^2}, \qquad i = 1, \cdots, s \tag{5}$$

is an over-all slope estimator.

Next we note that

$$\bar{y}_{st} = \sum_{t=1}^{L}(N_t/N)\bar{y}_t = \sum_{i=1}^{k}\bar{y}_{st}^i/k \tag{6}$$

where $\bar{y}_t$ is the mean of the n_t observations in the t-th stratum (similarly for $\bar{x}_{st}$) and finally that a conditionally unbiased estimator of Cov $(\bar{b}_{st}, \bar{x}_{st})$ is given by

$$(1 - n/N)\frac{1}{k(k-1)}\sum_{i=1}^{k}(\bar{x}_{st}^i - \bar{x}_{st})(b_{st}^i - \bar{b}_{st}).$$

Consequently, if $g = \bar{y}_{st} + \bar{b}_{st}(\mu_X - \bar{x}_{st})$ then $E(g) = \mu_Y -$ Cov $(\bar{b}_{st}, \bar{x}_{st})$ and therefore

$$T_{k(st)} = \bar{y}_{st} + \bar{b}_{st}(\mu_X - \bar{x}_{st}) + \left(1 - \frac{n}{N}\right)\frac{1}{k(k-1)}\sum_{i=1}^{k}(\bar{x}_{st}^i - \bar{x}_{st})(b_{st}^i - \bar{b}_{st}) \tag{7}$$

is a combined stratified unbiased estimator of μ_Y. Note that since $N_t = sn_t/k$, $k/s = n/N$. Nieto [1958] discussed the efficiency of the estimator (7) (for sampling with replacement) in detail.

Again the generalization to p auxiliary variates is straightforward.

As a specific illustration consider the case in which

$$b_{st}^{(i)} = \bar{y}_{st}^{(i)}/\bar{x}_{st}^{(i)} = r_{st}^{(i)}.$$

Then $T_{k(st)}$ reduces to

$$T_{k(st)} = \bar{r}_{st}\mu_X + \frac{Nk - n}{N(k-1)}(\bar{y}_{st} - \bar{r}_{st}\bar{x}_{st}). \tag{8}$$

In the special case that $N_t = \bar{N}$, $n_t = \bar{n}$ for all t and $k = \bar{n}$, $s = \bar{N}$ then

$$T_{k(st)} = \bar{r}_{st}\mu_X + \frac{(\bar{N} - 1)\bar{n}}{(\bar{n} - 1)\bar{N}}(\bar{y}_{st} - \bar{r}_{st}\bar{x}_{st}), \tag{9}$$

which is a generalized Hartley-Ross estimator.

Finally, we again consider an averaging of T_k over all possible splits of the sample into groups of size n_t/k, $t = 1, \cdots, L$. For this, the coefficient is taken in the form $b_{st}^{(i)} = r_{st}^{(i)} = \sum_{t=1}^{L} (N_t/N) r_t^{(i)}$ where $r_t^{(i)} = (k/n_t) \sum_{j=1}^{n_t/k} (y_j/x_j)$. Therefore,

$$\bar{r}_{st} = \sum_{i=1}^{k} r_{st}^{i} \Big/ k = \sum_{t=1}^{L} (N_t/N) \sum_{j=1}^{n_t} (y_j/x_j) \Big/ n_t = \sum_{t=1}^{L} (N_t/N) \bar{r}_t$$

and some algebraic reduction will show that $T_{k(st)}$ averaged over all possible splits is equal to

$$T_{k(st)}^{*} = \bar{r}_{st} \mu_X + (\bar{y}_{st} - \bar{r}_{st} \bar{x}_{st}) + \left(1 - \frac{n}{N}\right) \sum_{t=1}^{L} \frac{N_t^2}{N^2} \frac{(\bar{y}_t - \bar{r}_t \bar{x}_t)}{(n_t - 1)}, \tag{10}$$

which does not quite reduce to a form similar in appearance to Equation (8) and the Hartley-Ross estimator.

As before other selections of coefficients will yield other unbiased estimators.

5. MULTISTAGE SAMPLING

We consider a population with N primaries of equal size $\bar{M}$ and the following sampling scheme. First select n primaries from the N available with equal probability with or without replacement. Then select with equal probability one of the splits of each of the primaries into s groups of size $\bar{m}/k$. Then with equal probability and without replacement draw k of the groups so that the sample size is $\bar{m}$ in each selected primary.

Consider now the conditional distribution for a fixed set of primaries and a fixed split of the primaries into s groups each. Then by Section 4, Equation (11) is an unbiased estimator of $\bar{\bar{Y}}_n$, the population mean of the n selected primaries.

$$T_{k(M)} = \bar{\bar{y}} + \bar{b}(\mu_X - \bar{\bar{x}}) + \left(1 - \frac{\bar{m}}{\bar{M}}\right) \frac{1}{k(k-1)} \sum_{i=1}^{k} (b^i - \bar{b})(\bar{\bar{x}}^i - \bar{\bar{x}}) \tag{11}$$

where

$$\bar{\bar{y}}^i = (k/n\bar{m}) \sum_{t=1}^{n} \sum_{j=1}^{\bar{m}/k} y_{tj}^i, \qquad \bar{\bar{y}} = (1/k) \sum_{i=1}^{k} \bar{\bar{y}}^i = (1/n\bar{m}) \sum_{t=1}^{n} \sum_{j=1}^{\bar{m}} y_{tj}$$

and similarly for x. The coefficient $b^{(i)}$ is again arbitrary in form.

Finally, the expectation of $T_{k(M)}$ over all possible primary selections is the average of $\bar{\bar{Y}}_n$ over all possible primary selections; this is μ_Y and $T_{k(M)}$ is unbiased in multi-stage sampling.

Again the selection of the coefficients yields estimators of different types. For example, an unbiased ratio estimator of the Hartley-Ross type generalized to multistage sampling can be obtained.

6. VARIANCE ESTIMATION

It is interesting to notice that the same two-stage sampling scheme can be used to form an estimate of the variance of T_k . First assume a negligible n/N (or k/s) and a fixed set of uncorrelated groups. T_k can now be written

$$T_k = \bar{y} - \frac{1}{k(k-1)} \sum_{i \neq j}^{k} b^i(\bar{x}^j - \mu_X) \tag{12}$$

and its conditional variance can be expressed in terms of the variances and covariances of the components in (12). Since T_k is conditionally unbiased this variance has expectation equal to the over-all variance. Substitution of unbiased estimators for each of the terms of the variance (plus some terms of zero expectation) yields (13) as an unbiased estimator of the variance of T_k .

$$v(T_k) = T_k^2 - \frac{1}{k(k-1)} \sum_{i \neq j}^{k} \bar{y}^i \bar{y}^j. \tag{13}$$

Although this procedure is unbiased it can be subject to high sampling error, particularly for small k.

TABLE 1

A SIMPLE EXAMPLE OF THE ESTIMATORS

Sample Number	Pairs in Sample	$\bar{y}$	T_2			T_2^*
			Split 1	Split 2	Split 3	
1	$P_1P_2P_3P_4$	3.500	7.167	6.667	6.500	6.778
2	$P_1P_2P_3P_5$	5.250	8.917	8.250	7.917	8.361
3	$P_1P_2P_3P_6$	7.500	11.000	10.167	9.667	10.278
4	$P_1P_2P_4P_6$	8.750	11.917	10.250	9.917	10.694
5	$P_1P_2P_4P_5$	6.500	10.000	8.667	8.500	9.056
6	$P_2P_3P_4P_5$	7.500	8.167	7.667	7.500	7.778
7	$P_2P_3P_5P_6$	11.500	10.000	8.667	8.500	9.056
8	$P_2P_3P_4P_6$	9.750	9.417	8.750	8.417	8.861
9	$P_1P_3P_4P_5$	7.250	9.417	8.750	8.417	8.861
10	$P_1P_3P_4P_6$	9.500	11.000	10.167	9.500	10.222
11	$P_1P_3P_5P_6$	11.250	11.917	10.250	9.917	10.694
12	$P_3P_4P_5P_6$	13.500	7.167	6.667	6.500	6.778
13	$P_1P_4P_5P_6$	12.500	11.000	10.167	9.667	10.278
14	$P_2P_4P_5P_6$	12.750	8.917	8.250	7.917	8.361
15	$P_1P_2P_5P_6$	10.500	13.167	10.667	10.500	11.444

7. NUMERICAL ILLUSTRATION

To illustrate T_k a small population consisting of the six pairs $P_i = (y_i , x_i)$, $i = 1, 2, \cdots , 6$, with $y = x^2$ and $x_i = 0, 1, 2, \cdots , 5$ was completely examined Table 1 presents the values of $\bar{y}$, T_2 and T_2^* [using Equation (3)] for all possible samples of size four. For T_2, each of the possible samples was split into two groups of two in all possible ways, and the value of T_2 was computed for each. The three distinct values of T_2 for each sample are presented in the table. The numbering of the splits within a sample is of course arbitrary. It is readily verified that the average value of each of $\bar{y}$, T_2 and T_2^* is the population mean $\mu_Y = 9.167$. Furthermore, the exact population variances of $\bar{y}$, T_2 and T_2^* are 7.914, 2.281 and 1.886 respectively.

As a second example, the six pairs (y_i , x_i) were taken as follows: (0,2), (1, 3), (2, 5), (4, 9), (8, 14), (9, 15). All possible samples of size four were drawn and for each sample $\bar{y}$, y', $\hat{y}_b$, T_k (for all possible splits) and T_k^* were computed. A summary of the computations is presented in Table 2.

TABLE 2

ILLUSTRATION OF RELATIVE EFFICIENCIES

	Estimator				
	$\bar{y}$	y'	$\hat{y}_b$	T_k	T_k^*
Expectation	3.937	4.000	3.961	4.000	4.000
Bias	−0.063	0.000	−0.039	0.000	0.000
Variance	0.120	0.233	0.027	0.033	0.022
Mean Square Error	0.124	0.233	0.029	0.033	0.022

REFERENCES

Cochran, W. G., [1953]. *Sampling Techniques.* New York, John Wiley and Sons.

Goodman, L. A., and Hartley, H. O., [1958]. The precision of unbiased ratio-type estimators. *J. Amer. Stat. Assoc.* 53. 491–508.

Hartley, H. O., and Ross, A., [1954]. Unbiased ratio estimators. *Nature 174*, 270–271.

Nieto, J., [1958]. *Unbiased ratio estimators in stratified sampling.* Unpublished M.S. Thesis, Iowa State University, Ames, Iowa.

Robson, D. S., [1957]. Application of multivariate polykays to the theory of unbiased ratio-type estimation. *J. Amer. Stat. Assoc.* 52. 511–522.

A BIOMETRIC THEORY OF MIDDLE AND LONG DISTANCE TRACK RECORDS[1]

MALCOLM E. TURNER AND ELEANOR D. CAMPBELL
Medical College of Virginia, Richmond 19, Virginia, U. S. A.

1. *Introductory Remarks.*

It has been often observed that a plot of average velocity against distance or time for world track records suggests a continuous monotonic underlying relationship. Several attempts to discern the nature of this relationship empirically by seeking linearizing transformations have been made with varying degrees of success. The purposes of this paper are to describe an attempt to explain the observed curve by a biometric theory and to consider the problem of estimating unknown parameters occurring in the theory. The theory is applied to world record data of 1959 and records which are below par are pointed out. The development is based on the work of A. V. Hill [1927].

2. *Biometric Theory.*

Consider the situation in which a runner runs a given distance x in a certain time t at a constant speed $\dot{x}$, this constant speed not necessarily being the runner's maximum speed. We further define:

- y, the total energy (measured as volume of oxygen) excess over resting required to run the given distance with the given constant speed,
- $\dot{x}_0$, the greatest speed for which the energy (oxygen) required is not greater than the energy available for an indefinitely long period of time. Thus, at speed $\dot{x}_0$ the energy required is equal to the maximum amount of oxygen which can be brought into the body from the atmosphere,
- r, the maximum excess (over resting) rate at which oxygen can be brought into the body from the atmosphere,
- D, the maximum tolerable oxygen debt beyond which complete exhaustion is manifest.

The total excess energy y is a function of the constant speed $\dot{x}$ and the time of the race t. We assume that the rate of consumption of energy (oxygen) $\partial y/\partial t$ is a function of $\dot{x}$ but not of t. Thus,

[1] Read before the spring meeting of ENAR, April 1960, and based in part on a Ph.D. thesis supervised by Robert J. Monroe and H. L. Lucas, Jr., North Carolina State College, 1959.

$$\partial y/\partial t = f(\dot{x}) \text{ and } y = f(\dot{x})t.$$

The constant of integration is taken to be zero since no excess energy is required to run a race of zero duration.

The function $f(\dot{x})$ is of unknown form; however, we expect it to have at least the following characteristics:

(1) $f(\dot{x})$ is a non-decreasing function of $\dot{x}$.
(2) $f(\dot{x})$ has lower-order derivatives which vanish at the equilibrium velocity $\dot{x}_0$,
(3) $f(\dot{x}_0) = r$.

The first characteristic has been experimentally demonstrated by Hill [1927]. The second characteristic, although not experimentally demonstrated, seems plausible by analogy with other physiological equilibrium phenomena. It often happens that in the neighborhood of a point of equilibrium (here there is an equilibrium between the energy being consumed and the oxygen being brought into the cells) that compensatory physiological mechanisms come into play. It may be that circulatory changes or viscosity changes (due to temperature changes) provide such compensation in the present case. Whether or not these mechanisms are plausible, it seems appropriate to require characteristic (2) in the absence of experimental fact since one is hard pressed to find any physiological example not possessing this characteristic. Characteristic (3) is implied by the definition of $\dot{x}_0$ and r.

Perhaps the simplest function possessing these characteristics is the following:

$$f(\dot{x}) = r + b(\dot{x} - \dot{x}_0)^k, \qquad \dot{x} \geq \dot{x}_0\,. \tag{1}$$

Then we have

$$y = rt + b(\dot{x} - \dot{x}_0)^k t. \tag{2}$$

Equation (2) can be used to describe the experimental data shown in Figure 11 of Hill [1927]. Hill measured directly the excess oxygen required by an athlete to run for given periods of time at constant speeds. This experiment, of course, does not correspond to actual contest conditions, but does serve to illustrate the relation (2), which was not given by Hill. The meaning of the relation is clear. If a runner runs at his own asymptotic speed $\dot{x}_0$, then the excess energy required is obviously proportional to the time run where the proportionality constant is r; but, if he runs faster than $\dot{x}_0$, then the energy required will be increased monotonically with the divergence of his speed from the asymptotic speed.

Now, the speed cannot be increased indefinitely since there is an upper bound on energy available. This upper bound is given by

$$y_{\max} = rt + D. \tag{3}$$

For all but the shorter races the time lost in attaining optimum speed is negligible and it can be seen that the best time for a race can be obtained by maintaining a nearly constant speed throughout the race. It may be supposed that records are broken when a runner with a great amount of available energy is able to select that constant speed which equates the maximum available energy to the required energy. In this case

$$rt + D = rt + b(\dot{x} - \dot{x}_0)^k t\ , \tag{4}$$

which upon cancelling rt from both sides and rearranging gives an equation relating speed to time run. Thus,

$$\dot{x} = \dot{x}_0 + (bt/D)^{-1/k}. \tag{5}$$

The constants $\dot{x}_0$, D, b, and k are constants pertaining to the individual runner. D is the maximum oxygen debt the runner can tolerate and b and k relate to the mechanical efficiency of the runner's body, i.e., to the amount of work required to overcome the viscosity of the muscles. $\dot{x}_0$ is a function of the rate at which oxygen can be gotten to the tissues. Relation (5) could be used to study the progress of an individual runner in his period of training; however, our interest at present is in characterizing the world record curve. Although $\dot{x}_0$, D, b, and k are personal parameters, it may be supposed that, for world record holders, $\dot{x}_0$ and D approach the human maximum for all and that b and k are similarly close to the human minimum for all and that therefore it is appropriate to treat $\dot{x}_0$, D, b, and k as constant parameters for all record holders as a first approximation. Equation (5) is a special case of the single process law described by Turner [1959].

3. *Estimation Theory.*

For convenience we change the notation. Let:

$$Y = \dot{x}, \qquad \delta = -1/k,$$

$$\alpha = \dot{x}_0\ , \qquad X = t,$$

$$\beta = (b/D)^{-1/k} = (D/b)^{1/k}.$$

Then (5) may be rewritten:

$$Y = \alpha + \beta X^{\delta}. \tag{6}$$

Now, (6) cannot be expected to hold exactly for world records but it may be considered to be the limiting curve which is approached by the successive records in each event. Thus, for the actual observed records, we may write

$$Y = \alpha + \beta X^{\delta} \epsilon', \tag{7}$$

where $\epsilon' \leq 1$. Perhaps, the limiting curve (6) itself is not absolutely stable since the human species is possibly changing in its athletic capacities due to nutritional and genetic effects. However, these changes may be regarded to take place slowly as compared with the rate at which records are broken.

At any point in time we will assume as a first approximation that the distribution of world records below its asymptotic value can be described by the *truncated hyperbolic distribution:*[2]

$$f(\epsilon') = 1/\rho\epsilon', \qquad e^{-\rho} \leq \epsilon' \leq 1. \tag{8}$$

If we substract α from both sides of (7) and take logarithms we obtain the linear form:

$$\log (Y - \alpha) = \log \beta + \delta \log X + \epsilon, \tag{9}$$

where $\epsilon = \log \epsilon'$. Now it will be seen that the additive error ϵ follows the *rectangular distribution*:

$$f(\epsilon) = 1/\rho, \qquad -\rho \leq \epsilon \leq 0. \tag{10}$$

Thus, the problem of estimation is reduced to a well known one: to fit a straight line when the errors have a rectangular distribution. Maximum likelihood estimates are readily found by minimizing the maximum error (see Turner, [1960]). The procedure is as follows:

(a) Choose a trial estimate of α and then plot $\log (Y - \alpha)$ against $\log X$.

(b) The line with minimum maximum error will either (i) pass through one point such that all other points fall below and such that two points will have the same maximum error, or (ii) pass through two points such that all other points fall below and such that one point will have the maximum error.

(c) In case (i) the line will be parallel (and hence have the same slope) as a line passing through the two points with maximum error. The intercept is then found by passing a line with the same slope through the upper point. In case (ii) both slope and

[2] The truncated hyperbolic distribution was chosen as being, perhaps, the simplest distribution producing high contact on the right and having a finite upper limit. Since the upper limit is the object of estimation, the lower portions of the distribution are relatively irrelevant.

TABLE I

World Track Records and Estimated Limiting Values

Distance		Time			Speed in m./sec.	
x Metric	English	t Record*	$\hat{t}$ Estimated	$\hat{t} - t$ Difference	$\dot{x}$ Record	$\hat{\dot{x}}$ Estimated
400.0 m.		45.2″	45.2″	0.0″	8.850	8.850
402.3 m.	440 yd.	45.7″	45.6″	0.1″	8.804	8.822
800.0 m.		1′ 45.7″	1′ 44.3″	1.4″	7.569	7.670
804.7 m.	880 yd.	1′ 46.8″	1′ 45.1″	1.7″	7.534	7.657
1,000.0 m.		2′ 18.1″	2′ 15.5″	2.6″	7.241	7.380
1,500.0 m.		3′ 38.1″	3′ 35.4″	2.7″	6.878	6.964
1,609.4 m.	1 mi.	3′ 54.4″	3′ 53.1″	1.3″	6.866	6.904
2,000.0 m.		5′ 2.2″	4′ 58.0″	4.2″	6.618	6.711
3,000.0 m.		7′ 52.8″	7′ 47.2″	5.6″	6.345	6.421
3,218.7 m.	2 mi.	8′ 32.0″	8′ 24.9″	7.1″	6.287	6.375
4,828.0 m.	3 mi.	13′ 10.8″	13′ 4.9″	5.9″	6.105	6.151
5,000.0 m.		13′ 35.0″	13′ 34.7″	0.3″	6.135	6.137
9,656.0 m.	6 mi.	27′ 43.8″	27′ 29.0″	14.8″	5.804	5.856
10,000.0 m.		28′ 30.4″	28′ 30.4″	0.0″	5.847	5.847
Asymptotic Speed						5.103

*Source: *Information Please Almanac 1960* [1959].

intercept are found for the line passing through the two upper points. For either case the maximum error is readily ascertained. The choice of case (i) or (ii) is quickly made graphically and other determinations are made analytically.

(d) The above three steps are repeated for other choices of α and then, from the relation between minimum maximum error and α, one interpolates for the maximum likelihood estimate of α. Repeating the above three steps one final time with the optimum value of α leads to maximum likelihood estimates for β, δ, and ρ.

The above procedure illustrates a general theorem concerning minimum maximum error approximation by polynomial equations due to Vallée-Poussin [1911].

4. *Fit of Theory to World Track Records.*

The 1959 World records for all events between 400 and 10,000 meters (including English distances of 440 yards to six miles) were used to fit the theory. These data are given in Table I. Shorter distances and longer distances bring into play mechanisms unaccounted for in the theory, such as delay due to starting, muscle stiffness, blisters, nutritional requirements, etc.

Performing steps (a), (b), and (c) we obtain the following results:

$\hat{\alpha}$	$\hat{\rho}$ Minimized Maximum Error
5.08	0.03249
5.09	0.03190
5.10	0.03130
5.11	0.03143
5.12	0.03183

Parabolic interpolation between the middle three values in the above table yields the value $\hat{\alpha} = 5.103$. Refitting the straight line gives the estimates $\hat{\delta} = -0.4451$, $\hat{\beta} = 20.44$, and $\hat{\rho} = 0.03115$. We then have the estimated limiting curve.

$$\hat{x} = 5.103 + 20.44t^{-0.4451}, \tag{11}$$

where the estimated asymptotic value of the speed is $\hat{x}_0 = 5.103$. The estimated limiting speeds have been calculated according to equation

(11) and are shown in the last column of Table I beside the actual record speeds. Both record and estimated speeds are also shown in Figure 1. Although no very long races have been included in the analysis it is interesting to compare the speed for 30,000 *m.*, $\dot{x} = 5.26$, with the estimated asymptotic speed, $\hat{\dot{x}}_0 = 5.10$.

Estimated times $\hat{t}$ may be readily found by dividing the distance x by the estimated speeds $\hat{\dot{x}}$. These values are shown in column (4) of Table I and the differences between these estimated limiting times and the actual record times are given in column (5) of the same table. The extent of the agreement is almost startling. The largest discrepancy is seen to be just 14.8 seconds for the six-mile run. The time for the mile run appears to be relatively good, along with the 400, 5,000 and 10,000 meter races. With the exception of the mile run the "English" distances tend to lag behind the corresponding "metric" distances. Perhaps, this fact reflects nothing more than the popularity of the various distances, but one is tempted to speculate that large improvements in those races which lag behind may be forthcoming.

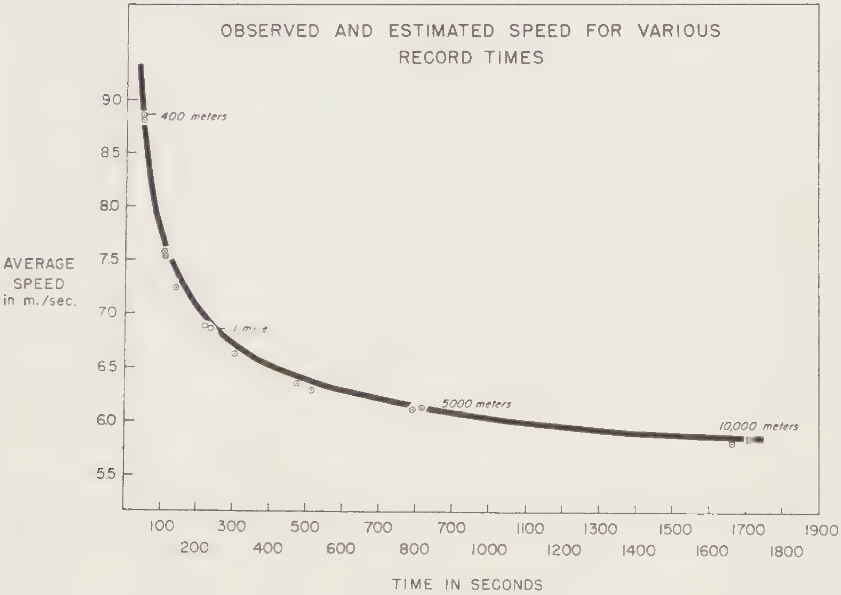


FIGURE 1

Observed and Estimated Speeds for Various Record Times

REFERENCES

Hill, A. V. [1927]. *Muscular Movement in Man.* McGraw-Hill Book Co., Inc., New York.

Information Please Almanac 1960. [1959]. McGraw-Hill Book Co., Inc., New York.

Turner, Malcolm E. [1959]. *The Single Process Law: A Study in Nonlinear Regression,* Dissertation No. 59-6571 (North Carolina State College). University Microfilms, Inc., Ann Arbor, Mich.

Turner, Malcolm E. [1960]. Note: On heuristic estimation methods. *Biometrics 16,* 299–300.

Vallée-Poussin, C. J. de la. [1911]. Sur la méthode de l'approximation minimum. *Annales de la Soc. Sc. de Bruxelles 35,* B: 1.

ADDENDUM

Since this analysis was completed Sweden's Dan Waern lowered the 1000 m. record to 2′ 17.8″ and Australia's Herb Elliott lowered the 1500 m. record to 3′ 36.0″. These are both above the estimated optimum times for these distances and hence the analysis is not affected. However, the fit is substantially improved.

SMALL SAMPLE BEHAVIOR OF SLOPE ESTIMATORS IN A LINEAR FUNCTIONAL RELATION[1]

MARTIN DORFF
University of Maine, Orono, Maine, U. S. A.
AND
JOHN GURLAND[2]
Mathematics Research Center, U. S. Army, University of Wisconsin
Madison, Wisconsin, U. S. A.

1. INTRODUCTION

This paper is concerned with estimating the slope parameter in a linear functional relation between two variables which are not observable since both are subject to error. Such problems occur in many contexts. As a matter of fact, in virtually all applications of fitting a functional relation errors of observation occur in all variables. Very frequently we can neglect errors in the independent variables in comparison with the errors in the dependent variable, but clearly we cannot always do so. In particular, it is often desirable to use a relatively inexpensive technique in place of a very costly one, where the results given by the two, apart from errors, are linearly related. For example, the evaluation of rocket grains by static test techniques is relatively inexpensive compared to the procedure of utilizing dynamic or flight testing.

The following linear relation is assumed between the variables X, Y.

$$Y = \alpha + \beta X.$$

Neither X nor Y can be observed, but only x and y, where

$$x = X + e$$
$$y = Y + f;$$

e and f are independent random variables with expectation zero representing errors of observations; X and Y are the expected values of x and y respectively. When for each $X_i(Y_i)$ we have a single observation $x_i(y_i)$, we shall say that we have unreplicated observations on that variable; otherwise we have replicated observations.

[1] Preparation of this paper for publication was supported in part by contract No. DA-11-022-ORD-2059; research supported in part by Air Force Office of Scientific Research, Contract No. AF 49(639)43.

[2] On leave from Iowa State University of Science and Technology, Ames, Iowa, U. S. A.

In a previous paper [3] we considered the large sample properties of various estimators of the parameters α, β. In the present paper we consider the small sample behavior of these slope estimators in order to compare them. First however, we take up the small sample behavior of slope estimators which do not require replicated observations; such estimators were discussed only briefly in our earlier paper.

We have chosen to examine the bias and the mean square error of the various estimators, partly because of the innate interest which these properties hold and partly because they seemed most amenable to attack. In all the cases considered the square of the bias is small compared with the mean square error; consequently our findings concerning the latter are essentially valid for the variance as well.

2. ESTIMATORS WHICH DO NOT REQUIRE REPLICATION

The problem of estimating the parameters of a linear functional relation when both variables are in error and there is no replication has long been regarded as an intractable one. In 1940 Wald [10] suggested that, when the number of observations n is even, one can divide the observations into two equal groups and take as an estimator of β

$$b_W = \left(\sum_{\frac{1}{2}n+1}^{n} y_i - \sum_{1}^{\frac{1}{2}n} y_i\right) \Big/ \left(\sum_{\frac{1}{2}n+1}^{n} x_i - \sum_{1}^{\frac{1}{2}n} x_i\right).$$

He showed that b_W is a consistent estimator of β provided that the partition of observations into two groups can be carried out independently of the errors and provided that limit inferior of

$$|\,(X_1 + \cdots + X_{\frac{1}{2}n}) - (X_{\frac{1}{2}n+1} + \cdots + X_n)\,|/n$$

is positive. If the errors in the x_i are small enough so that partitioning the observations according to the magnitudes of the x_i yields the same two groups as obtained by partitioning the observations according to the magnitudes of the X_i, the first condition is satisfied. The second condition guarantees that the expectation of the denominator does not vanish.

Nair and Shrivastava [8], Bartlett [1], and others [4], [9], have shown that, for certain spacings of the X_i, more efficient estimators of β can be achieved by dividing the observations into three groups and omitting the middle group. In particular, Bartlett suggested that, when the number of observations is divisible by three, one should take as an estimator of β

$$b_B = \left(\sum_{\frac{2}{3}n+1}^{n} y_i - \sum_{1}^{\frac{1}{3}n} y_i\right) \Big/ \left(\sum_{\frac{2}{3}n+1}^{n} x_i - \sum_{1}^{\frac{1}{3}n} x_i\right).$$

This is a consistent estimator of β under the obvious modification of

Wald's conditions. Bartlett showed that b_B is more efficient than b_W when the X_i are equally spaced.

Housner and Brennan [6] suggest that the estimator

$$b_H = \sum (i - \bar{\imath})y_i / \sum (i - \bar{\imath})x_i ,$$

where i now denotes the order number when the x's are ordered according to increasing magnitude, will often be more efficient than any of the foregoing. In fact they give empirical results to substantiate the belief that b_H is more efficient than b_W when the X_i are equally-spaced. It should be realized, however, that consistency of this estimator rests upon more stringent assumptions than those made by Wald and the other writers mentioned above. One must now assume the entire ordering according to the magnitudes of the x_i to be identical with the ordering according to the X_i .

It is clear that all of the estimators mentioned above are members of the class $b_L = \sum w_i y_i / \sum w_i x_i$, where i denotes the order number when the x's are ordered according to increasing magnitude and $\sum w_i = 0$. One requires this restriction on the w_i in order to ensure that b_L be invariant with respect to translations of the coordinate axes. In our investigation of b_L we shall make the assumption that the ordering according to the magnitude of the x_i is identical with the ordering according to the X_i . If that is so, the w_i are simply constants, and b_L is the ratio of two linear forms in x_i and y_i . In some fields of research such an assumption might be unrealistic, but in the physical sciences it would ordinarily be regarded as quite reasonable. In any event, occurence of situations where the assumption is reasonable seem sufficiently frequent to justify a detailed investigation.

We propose now to examine the bias and mean square error of b_L as they relate, in particular, to the choice of the w_i . We first set down the model and the assumptions which underlie the discussion in Sections 3, 4, and 5 below. We suppose that

$$Y_i = \alpha + \beta X_i , \qquad i = 1, 2, \cdots , n,$$

where α and β are unknown parameters while the X_i and Y_i are unknown constants. The X_i are assumed to be ordered according to increasing magnitude. We observe x_i and y_i , where

$$x_i = X_i + e_i ,$$

$$y_i = Y_i + f_i .$$

The e_i and f_i are random variables, representing errors of observation, such that

(1) e_i and e_j are independent if $i \neq j$,
(2) f_i and f_j are independent if $i \neq j$,
(3) e_i and f_j are independent for all i and j,
(4) $E(f_i) = 0$ and $E(f_i^2) = \sigma_f^2 = \nu_2$.
(5) Let $c = \min_i \mid X_{i+1} - X_i \mid$. Then Prob $\{\mid e_i \mid \geq \frac{1}{2}c\} = 0$. That is to say, the e_i have finite range, extending from $-\frac{1}{2}c$ to $\frac{1}{2}c$. This is the assumption which ensures that the ordering according to the x_i is identical with the ordering according to the X_i .
(6) $E(e_i) = 0$ and $E(e_i^2) = \sigma_e^2 = \mu_2$, $E(e_i^4) = \mu_4$, $E(e_i^6) = \mu_6$. All odd moments of the e_i are zero.
(7) $\sum w_i X_i \neq 0$. It is easily verified that for each of the estimators considered this is the case.

It should be stressed that the e_i and f_i represent errors of observation only since the X_i and Y_i are constants, otherwise e_i and f_i would obviously not be independent. As for assumption (5), it also ensures that the estimators are consistent, since $\sum w_i = 0$. In practice it is conceivable that one might wish to relax assumption (5) slightly to permit a distribution with infinite range but with most of the probability concentrated in $(-c/2, c/2)$. The extent to which one might relax this assumption is not considered in this paper but will be investigated subsequently.

3. BIAS AND MEAN SQUARE ERROR OF b_L.

The bias, B_L of b_L , may be obtained in the following way:

$$E(b_L) = E(\sum w_i y_i) E\left(\frac{1}{\sum w_i x_i}\right).$$

Now

$$\frac{1}{\sum w_i x_i} = \frac{1}{P + \sum w_i e_i}$$
$$= \frac{1}{P} \sum_{k=0}^{t} (-1)^k P^{-k} (\sum_i w_i e_i)^k + \frac{(-1)^{t+1}}{P^{t+2}} \frac{(\sum w_i e_i)^{t+1}}{1 + \dfrac{\sum w_i e_i}{P}},$$

where $P = \sum w_i X_i$.
Consequently,

$$E\left(\frac{1}{\sum w_i x_i}\right) = \frac{1}{P} \sum_{k=0}^{t} (-1)^k P^{-k} E(\sum w_i e_i)^k + D,$$

where

$$D = \frac{(-1)^{t+1}}{P^{t+2}} E\left\{(\sum w_i e_i)^{t+1} \Big/ \left[1 + \frac{\sum w_i e_i}{P}\right]\right\}.$$

It has proved sufficiently accurate for the purpose of this investigation to take $t = 5$. Then

$$E\left(\frac{1}{\sum w_i x_i}\right) = \frac{1}{P} + \frac{1}{P^3} E(\sum w_i e_i)^2 + \frac{1}{P^5} E(\sum w_i e_i)^4 + D,$$

where

$$D = \frac{1}{P^7} E\left[(\sum w_i e_i)^6 \Big/ \left(1 + \frac{\sum w_i e_i}{P}\right)\right].$$

Since $E(\sum w_i y_i) = \beta P$, we have

$$E(b_L) = \beta\left[1 + \frac{1}{P^2} E(\sum w_i e_i)^2 + \frac{1}{P^4} E(\sum w_i e_i)^4\right] + \beta P\ D.$$

Therefore

$$B_L = \beta\left[\frac{1}{P^2} \mu_2 \sum w_i^2 + \frac{1}{P^4} (\mu_4 \sum w_i^4 + 3\mu_2^2 \sum_{i \neq j}\sum w_i^2 w_j^2)\right] + \beta P\ D.$$

In all of the cases studied it is easily verified that

$$-1 < \sum w_i e_i/P < 1.$$

Consequently, for the cases we want to investigate it is possible to find simple bounds for D. In fact,

$$0 < D < \frac{1}{2P^7}\left[\frac{1}{1 - \frac{\frac{1}{2} \sum | w_i | c}{P}} + 1\right] E(\sum w_i e_i)^6.$$

To see this, one need only realize that for negative values of $\sum w_i e_i$

$$0 \leq \sum (w_i e_i)^6 \Big/ \left(1 + \frac{\sum w_i e_i}{P}\right) \leq \sum (w_i e_i)^6 \Big/ \left(1 - \frac{\frac{1}{2}c \sum | w_i |}{P}\right),$$

while for positive values of $\sum w_i e_i$

$$0 \leq \sum (w_i e_i)^6 \Big/ \left(1 + \frac{\sum w_i e_i}{P}\right) \leq \sum (w_i e_i)^6.$$

Consequently,

$$0 \leq E\left[\frac{\sum (w_i e_i)^6}{1 + \frac{\sum w_i e_i}{P}}\right] \leq \left[\frac{1}{1 - \frac{\frac{1}{2}c \sum | w_i |}{P}}\right] \tfrac{1}{2} E(\sum w_i e_i)^6$$

$$+ \tfrac{1}{2} E(\sum w_i e_i)^6.$$

Therefore

$$0 < D < \frac{1}{2P^7}\left[\frac{1}{1 - \frac{\frac{1}{2}\sum |w_i| c}{P}} + 1\right]$$

$$\cdot\left(\mu_6 \sum w_i^6 + 15\mu_4\mu_2 \sum_{i \neq j}\sum w_i^4 w_j^2 + 15\mu_2^3 \sum \sum_{i\neq j\neq k} \sum w_i^2 w_j^2 w_k^2\right).$$

Thus we have the means for approximating the bias to any described degree of accuracy, and furthermore, we have a way of assessing the accuracy of any approximation.

It is possible to obtain the mean square error, M_L of b_L, in essentially the same way as B_L was obtained. The result is

$$M_L = \beta^2\left[\frac{1}{P^2}\mu_2 \sum w_i^2 + \frac{3}{P^4}\left(\mu_4 \sum w_i^4 + 3\mu_2^2 \sum_{i\neq j}\sum w_i^2 w_j^2\right)\right]$$

$$+ \lambda \frac{1}{P^2}\left[\mu_2 \sum w_i^2 + \frac{3}{P^4}\mu_2^2\left(\sum w_i^2\right)^2\right.$$

$$\left. + \frac{5}{P^6}\sum w_i^2\left(\mu_4 \sum w_i^4 + 3\mu_2^2 \sum_{i\neq j}\sum w_i^2 w_j^2\right)\right]$$

$$+ \Delta, \quad \text{where} \quad \lambda = \nu_2/\mu_2$$

and

$$0 < \Delta < \left\{\beta^2 \frac{7}{2}\left(1 - \frac{\frac{1}{2}c \sum |w_i|}{P}\right)^{2} + \frac{5}{2} - \left(1 + \frac{\frac{1}{2}c\sum |w_i|}{P}\right)^{-1}\right.$$

$$\left. + \frac{\lambda\mu_2 \sum w_i^2}{P^2}\left[\frac{7}{2}\left(1 - \frac{\frac{1}{2}c\sum |w_i|}{P}\right)^{-2} + \frac{7}{2}\right]\right\}$$

$$\cdot\left\{\mu_6 \sum w_i^6 + 15\mu_4\mu_2 \sum_{i\neq j}\sum w_i^4 w_j^2 + 15\mu_2^3 \sum\sum_{i\neq j\neq k}\sum w_i^2 w_j^2 w_k^2\right\}.$$

4. OPTIMAL CHOICE OF WEIGHTS.

We should like now to consider two possible sets of w_i; namely,

1. The set which minimizes the bias B_L, and
2. The set which minimizes the mean square error M_L.

In general, of course, the two sets differ. The exact specifications of either set does not appear possible in view of the complexity of the expressions for B_L and M_L. On the other hand it is possible to obtain very good approximate specifications.

If one chooses the w_i according to the scheme of Wald, or Bartlett, or Housner and Brennan, $\max_{e_i} | \sum w_i e_i |/P$ turns out to be considerably less than unity, even for values of n as small as four; this is a consequence of assumption (5) of Section 2. This leads one to conjecture that max $|\sum w_i c_i|/P$ is small for any reasonable choice of the w_i—in particular, one would expect it to be small for good choices, like that which minimizes the bias or the mean square error. If this is the case, one would expect the bias and the mean square error to be determined primarily by the term of order P^{-2} in the expression for each; consequently, minimizing that term should approximately minimize the bias or the mean square error, as the case may be. Now this term in the expression for the bias is $\mu_2 \beta P^{-2} \sum w_i^2$, whereas the corresponding term in the expression for the mean square error is $\mu_2(\beta^2 + \lambda)P^{-2} \sum w_i^2$. Obviously both of these terms are minimized by the same set of w_i, which is easily shown to be given by $w_i = (X_i - \bar{X})C$, where C is any arbitrary constant unequal to 0.

Up to this point we have proceeded heuristically; we should now like to show with the aid of some examples that retention of terms in P^{-4} makes only a trifling difference in the weights and far less difference in the bias itself.

As the first example, we consider four points: $X_1 = -3c/2$, $X_2 = -c/2$, $X_3 = c/2$, and $X_4 = 3c/2$. We take $\mu_2 = \frac{1}{3}(c/2)^2$ and $\mu_4 = \frac{1}{5}(c/2)^4$ which are the moments of the rectangular distribution on the interval $(-c/2, c/2)$, which is the maximum possible range consonant with the assumptions of Section 2. We might equally well consider any other admissible distribution on this finite range; the same technique would apply; and the results would be substantially the same. The condition above gives as the weights $w_1 = -3$, $w_2 = -1$, $w_3 = +1$, $w_4 = +3$ (or any multiple of these, of course). Minimization of

$$B_L \doteq \beta\left[\frac{1}{P^2}\mu_2 \sum w_i^2 + \frac{1}{P^4}\left(\mu_4 \sum w_i^4 + 3\mu_2^2 \sum\sum w_i^2 w_j^2\right)\right]$$

gives for the w_i

$$w_1 = -3.044, \qquad w_2 = -1, \qquad w_3 = +1, \qquad w_4 = +3.044.$$

This introduces a relative change in the bias $\Delta B_L/B_L$ of 9×10^{-5}.

As a second example, we consider the four points $X_1 = -5c/2$, $X_2 = -c/2$, $X_3 = c/2$, and $X_4 = 5c/2$, with μ_2 and μ_4 just as in the previous example. The condition gives

$$w_1 = -5, \qquad w_2 = -1, \qquad w_3 = +1, \qquad w_4 = +5.$$

Minimization of

$$B_L \doteq \beta\left[\frac{1}{P^2}\mu_2 \sum w_i^2 + \frac{1}{P^4}\left(\mu_4 \sum w_i^4 + 3\mu_2^2 \sum\sum w_i^2 w_j^2\right)\right]$$

gives for the w_i

$$w_1 = -5.034, \qquad w_2 = -1, \qquad w_3 = +1, \qquad w_4 = 5.034.$$

This introduces a relative change in the bias $\Delta B_L/B_L = 3 \times 10^{-6}$.

It will be observed that the perturbations produced both in the weights and in the bias by the term in P^{-4} are smaller in the second example than in the first. That is exactly what one would expect, for although the spread of the X_i has increased in the second example, the range of the e_i has not changed. This has the effect of decreasing max $\sum w_i e_i/P$, thereby decreasing the importance of the term in P^{-4} relative to that of the term in P^{-2}. One might examine the mean square error in similar fashion. It is clear that the approximate minimization of bias and mean square error given by choosing $w_i = X_i - \bar{X}$ is a very good approximation, and the estimator b_O which employs these weights can for all practical purposes, be regarded as having minimum bias and minimum means square error. We shall refer to this choice of weights as optimal and to the estimator

$$[\sum (X_i - \bar{X})y_i]/[\sum (X_i - \bar{X})x_i]$$

as the optimal (ratio-of-linear-forms) estimator.

5. COMPARISON OF VARIOUS WELL-KNOWN ESTIMATORS WITH THE OPTIMAL ESTIMATOR.

In general the spacing of the X_i is unknown, and the optimal estimator is therefore not obtainable. It is customary to use the weights of Wald, of Bartlett, or of Housner and Brennan. The corresponding estimators will henceforth be denoted by b_W, b_B, b_H respectively, and the corresponding biases and mean square errors will also carry these subscripts. For example B_B denotes the bias of Bartlett's estimator.

When the X_i are uniformly spaced, b_H and b_O are identical, which is to say that the Housner-Brennan estimator is optimal for this spacing. For any other spacing of the X_i neither b_H, b_B, nor b_W is optimal; the question is, how far do they depart from optimality.

To answer this question we have studied the bias and the mean square error of these estimators for $n = 6$, 12 and for various types of spacings of the X_i, which may be classified as follows:

1. Symmetric spacing: $\cdots, -5^\omega p, -3^\omega p, -1^\omega p, 1^\omega p, 3^\omega p, 5^\omega p, \cdots$ for $\omega \geq 0$
2. Symmetric spacing: $-1^\omega p, -2^\omega p, -3^\omega p, \cdots, 3^\omega p, 2^\omega p, 1^\omega p$ for $\omega < 0$
3. Asymmetric spacing: $0, 2(1^\omega)p, 2(2^\omega)p, 2(3^\omega)p, \cdots$ for $\omega > 1$ and $0 < \omega < 1$
4. Asymmetric spacing: $-1^\omega p, -2^\omega p, -3^\omega p, \cdots$ for $\omega < 0$.

We have taken the values of p to make the ranges of the selected X_i values approximately the same, ensuring at the same time that condition (5) of Section 2 holds. In addition we have taken

$$\mu_2 = \frac{1}{3}\left(\frac{c}{2}\right)^2, \qquad \mu_4 = \frac{1}{5}\left(\frac{c}{2}\right)^4, \qquad \mu_6 = \frac{1}{7}\left(\frac{c}{2}\right)^6,$$

which coincide with the second, fourth, and sixth moments of a rectangular distribution having the maximum range consonant with the assumptions of Section 2. As indicated in Section 4 there is no serious loss of generality in confining our attention to a rectangular distribution of errors.

Tables 1 to 5 present, for $n = 6$ and $n = 12$,

1. The bias of the optimal estimator.
2. The ratio of the bias of b_W, b_B, and b_H to the bias of the optimal estimator.
3. The ratio of the mean square error of b_W, b_B, and b_H to that of the optimal estimator.

The values cited in the tables are correct to within one unit in the last place given there. Results for other values of n can be found in Dorff [2].

6. LEAST-SQUARES-TYPE ESTIMATOR

This section deals with $b_Q = [\sum(x_i - \bar{x})y_i]/[\sum (x_i - \bar{x})x_i$, which is the estimator one would obtain if he minimized the sum of squares of deviations in the vertical direction, simply ignoring the fact that the x_i are random variables. It is well-known that b_Q is not a consistent

TABLE 1

Maximum Value of B_Q/β.

n	Values of ω 2.0	 1.0	 0.5	 −0.5	 −2.0
			Symmetric Spacing		
6	.0002	.0048	.0012	.0004	.0008
12	.0000	.0006	.0001	.0001	.0000
			Asymmetric Spacing		
6	.0002	—[3]	.0014	.0006	.0000
12	.0000	—	.0002	.0000	.0000

[3] For $\omega = 1.0$ the pattern of spacing presented would be symmetric; consequently no value is cited here.

TABLE 2

Bias Ratio for Symmetric Spacing

Ratio	Values of ω				
	2.0	1.0	0.5	−0.5	−2.0
	$n = 6$				
B_H/B_O	1.06	1.00	1.04	1.08	1.10
B_W/B_O	1.73	1.30	1.09	1.06	1.74
B_B/B_O	1.22	1.10	1.14	1.25	1.38
B_Q/B_O	−3.00	−2.97	−2.90	−2.99	−2.99
	$n = 12$				
B_H/B_O	1.06	1.00	1.04	1.06	1.47
B_W/B_O	1.78	1.32	1.11	1.11	2.93
B_B/B_O	1.28	1.12	1.15	1.26	2.13
B_Q/B_O	−9.00	−9.00	−9.00	−9.00	−9.00

TABLE 3

Bias Ratio for Asymmetric Spacing

Ratio	Values of ω			
	2.0	0.5	−0.5	−2.0
	$n = 6$			
B_H/B_O	1.08	1.10	1.17	1.62
B_W/B_O	1.41	1.58	1.69	2.81
B_B/B_O	1.19	1.27	1.34	2.03
B_Q/B_O	−3.00	−3.00	−3.00	−3.00
	$n = 12$			
B_H/B_O	1.08	1.09	1.32	2.51
B_W/B_O	1.43	1.56	2.08	5.24
B_B/B_O	1.20	1.27	1.61	3.66
B_Q/B_O	−9.00	−9.00	−9.00	−9.00

TABLE 4

Mean-Square-Error Ratio for Symmetric Spacing

Ratio	Values of ω 2.0	1.0	0.5	−0.5	−2.0
			$n = 6$		
M_H/M_O	1.06	1.00	1.04	1.08	1.10
M_W/M_O	1.73	1.31	1.09	1.06	1.74
M_B/M_O	1.22	1.10	1.14	1.26	1.38
M_Q/M_O	0.99	—[4]	0.99	0.99	1.00
			$n = 12$		
M_H/M_O	1.06	1.00	1.04	1.07	1.47
M_W/M_O	1.78	1.32	1.11	1.11	2.93
M_B/M_O	1.28	1.12	1.15	1.26	2.13
M_Q/M_O	1.00	1.01	1.00	1.00	1.00

[4] Because of slow convergence of our approximation techniques for this combination of parameters, we have no reliable value to cite here.

TABLE 5

Mean-Square-Error for Asymmetric Spacing

Ratio	Values of ω 2.0	0.5	−0.5	−2.0
		$n = 6$		
M_H/M_O	1.09	1.11	1.17	1.62
M_W/M_O	1.41	1.58	1.69	2.81
M_B/M_O	1.19	1.27	1.34	2.03
M_Q/M_O	1.00	0.98	1.00	1.00
		$n = 12$		
M_H/M_O	1.08	1.09	1.32	2.51
M_W/M_O	1.43	1.56	2.08	5.24
M_B/M_O	1.20	1.27	1.61	3.66
M_Q/M_O	1.00	1.00	1.00	1.00

estimator of β; however, it is conceivable that it might have desirable small-sample properties. We shall now investigate the bias and mean square error of b_Q, making the same assumptions as in Section 2. Tables 2–5 cite values of B_Q/B_O and M_Q/M_O correct to within one unit in the last place given there. It will be observed that B_Q/B_O is approximately $-(n - 3)$ while M_Q/M_O is nearly unity.

7. DISCUSSION OF ESTIMATORS WHICH DO NOT REQUIRE REPLICATION

We should like now to compare the estimators discussed in Section 2 with a view toward making some practical suggestions for workers who are actually concerned with fitting straight-line relations.

Generally speaking b_B is clearly superior to b_W but there are some situations when such is not the case; we found that b_W had lower bias and lower mean square error than b_B when the X_i are symmetrically spaced, with $\omega = 0.5$ or $\omega = -0.5$. However, in almost every case investigated, b_H proved superior to b_W and to b_B. This, of course, is hardly surprising. The use of b_H presupposes through the ability to order all the observations more information available to the experimenter than that of b_W or b_B, and estimators based upon more complete information are typically more efficient than those based upon partial information. We did, nevertheless, find instances where b_W proved superior to b_B and b_H. One of these is $n = 6$, $\omega = -0.5$, symmetric spacing. For asymmetric spacings corresponding to these values of n and ω, b_H was found decidedly superior to the others.

As the comparison of b_H and b_Q is slightly more involved, we should like to first compare b_O with b_Q. The conclusion seems inescapable that when b_O can be obtained it is preferable to b_Q for two reasons:

1) b_O has much smaller bias than b_Q, while its mean square error is essentially no greater.
2) b_O is a consistent estimator of β, whereas b_Q is not.

The difficulty, of course, is that b_O cannot usually be obtained, inasmuch as the spacing of the X_i is unknown.

However it appears that b_H is a reasonable substitute for b_O whenever the X_i possess no marked skewness and are not bunched excessively. When such is the case we see that neither B_H/B_O nor M_H/M_Q will greatly exceed unity, while B_H/B_Q will be approximately $-1/(n - 3)$. Thus, when using b_H in preference to b_Q, one stands to do much better with respect to bias and very little worse with respect to mean square error.

One could roughly assess the skewness or bunchiness of his data by plotting the x_i versus i on log-log paper; the slope of the line which best fits the points is then ω. Ordinarily this should not be necessary; in most cases one could decide whether there was evidence of skewness or bunchiness by merely looking at the data.

8. ESTIMATORS WHICH REQUIRE REPLICATION.

We consider now various estimators of β suitable when replicated observations are available. We suppose that

$$Y_i = \alpha + \beta X_i \,, \qquad i = 1, 2, \cdots, n,$$

just as in Section 2, but now we have observations

$$x_{it} = X_i + e_{it} \,,$$

$$y_{it} = Y_i + f_{it} \,,$$

where $t = 1, 2, \cdots, r (r > 1)$;
e_{it} and f_{it} are random variables, representing errors of observation, such that

1) e_{it} and e_{ju} are independent unless $i = j$ and $t = u$.
2) f_{it} and f_{ju} are independent unless $i = j$ and $t = u$.
3) e_{it} and f_{ju} are independent for all i, j, t, u.
4) Let $c = \min_i | X_{i+1} - X_i |$. Then Prob $\{| e_{it} | \geq c/2\} = 0$.
5) $E(e_{it}^2) = \mu_2 \,, E(e_{it}^4) = \mu_4 \,, E(e_{it}^6) = \mu_6$. All odd moments of the e_{it} are zero.
6) Prob $\{| f_{it} | \geq \frac{1}{2}h\beta c\} = 0$. That is to say, we now restrict f_{it} to a finite range $(-\frac{1}{2}h\beta c, \frac{1}{2}h\beta c)$; the symbol h is defined by this relation.
7) $E(f_{it}^2) = \nu_2 = h^2\beta^2\mu_2 \,, E(f_{it}^4) = \nu_4 = h^4\beta^4\mu_4 \,, E(f_{it}^6) = \nu_6 = h^6\beta^6\mu_6$.

   All odd moments of the f_{it} are zero. (These assumptions would be satisfied if the distribution of the f_{it} differed from that of the e_{it} only in its range.)

In our previous paper [3] we considered the following estimators of β, all of which are consistent provided that their denominators have non-vanishing expectations:

$$b_1 = B_{xy}/(B_{xx} - W_{xx}), \;\; b_2 = (B_{yy} - W_{yy})/B_{xy} \,, \;\; b_3 = \sqrt{\frac{B_{yy} - W_{yy}}{B_{xx} - W_{xx}}} \operatorname{sgn} \beta$$

where

$$B_{xx} = \frac{r}{n - 1} \sum (x_i - x_{..})^2, \qquad W_{xx} = \frac{1}{n(r - 1)} \sum \sum (x_{it} - x_{i.})^2,$$

$$B_{yy} = \frac{r}{n-1} \sum (y_{i.} - y_{..})^2, \quad W_{yy} = \frac{1}{n(r-1)} \sum \sum (y_{it} - y_{i.})^2,$$

$$B_{xy} = \frac{r}{n-1} \sum (x_{i.} - x_{..})(y_{i.} - y_{..}),$$

and sgn $\beta = +1$ if $\beta > 0$
-1 if $\beta < 0$.

We have applied the techniques of Section 3 to the study of the bias and mean square error of these three estimators for comparison with the bias and mean square error of $b_O = [\sum (X_i - \bar{X})y_i]/[\sum (X_i - \bar{X})x_i]$. We found that b_1, b_2, b_3 and b_O have approximately the same mean square error but that their biases can differ greatly. In fact, when the errors are sufficiently small $B_1/B_O \doteq 2$, $B_2/B_O \doteq 1 - h^2$, $B_3/B_O \doteq (3 - h^2)/2$. Thus, the absolute bias of b_1 is smaller than that of b_2 whenever $h > \sqrt{3}$; the absolute bias of b_3 is smaller than that of b_1 and b_2 whenever $\sqrt{5/3} < h < \sqrt{7}$. Now $h^2 = \nu_2/\mu_2/\beta^2$, and since ν_2/μ_2 is ordinarily unknown, we are unable to decide whether $\sqrt{5/3} < h < \sqrt{7}$ or not. This would seem to rule out b_3 as a useful estimator of β. On the other hand in many problems it might well be possible to decide whether $h > \sqrt{3}$ or not. Inasmuch as b_1 and b_2 have approximately the same mean square error one could then use the bias as a criterion for choosing between the two: i.e.,

$$\text{When} \quad \beta^2 < \frac{\nu_2/\mu_2}{3}, \qquad \text{use} \quad b_1 ;$$

$$\text{When} \quad \beta^2 > \frac{\nu_2/\mu_2}{3} \qquad \text{use} \quad b_2 .$$

It is interesting to note that analysis of the asymptotic variance of b_1 and b_2 in our previous paper [3] led us to recommend that

$$\text{when} \quad \beta^2 < \nu_2/\mu_2, \qquad \text{use} \quad b_1,$$

$$\text{when} \quad \beta^2 > \nu_2/\mu_2, \qquad \text{use} \quad b_2,$$

so that the large sample approach based on variance and the small sample approach based on bias lead to somewhat different conclusions.

For values of h larger than $\sqrt{2}$, b_O would be preferable to b_1, but unfortunately b_O is not usually obtainable. We could, of course, use b_H, b_W, or b_B in lieu of b_1 at some sacrifice of efficiency, but with a corresponding reduction of the magnitude of the bias. If one knew something about the spacing of the X_i, this might be a desirable way to obtain an estimate of β, but if nothing is known about the X_i, it seems advisable to use either b_1 or b_2. In any case the use of b_1 or b_2 does not require any assumption about the possibility of ordering the observed values correctly in contradistinction to the estimators b_H, b_W, b_B, or b_O.

9. CORRELATED ERRORS

In some investigations it would not be reasonable to assume the errors e_i and f_i to be independent; one should properly take account of correlation in such cases. The correlation of errors greatly increases the complexity of the problem and an examination of small sample properties for such situations has not been included here.

We have, however, examined in [2], [3] the asymptotic variance for consistent estimators of β when the errors are correlated. In this case, the estimators b_1, b_2, and b_3 are modified as follows in order to retain the property of consistency:

$$b_1 = (B_{xy} - W_{xy})/(B_{xx} - W_{xx}), \qquad b_2 = (B_{yy} - W_{yy})/(B_{xy} - W_{xy}),$$

$$b_3 = \sqrt{(B_{yy} - W_{yy})/(B_{xx} - W_{xx})}\ \operatorname{sgn} \beta.$$

Essentially the same conclusions were obtained as in the case of independent errors:

1. When $\beta^2 < \nu_2/\mu_2$, use b_1,
2. When $\beta^2 > \nu_2/\mu_2$, use b_2,
3. b_3 is advantageous for only a narrow range of values of the parameters and is therefore not likely to be of much interest.

10. CONCLUSIONS

In this paper we have considered the problem of estimating the slope of a linear functional relation when both variables are in error.

When replicated observations are made a number of estimators are available, all of which have approximately the same asymptotic variance and mean square error, but study of the bias enables us to choose among them when we have partial information concerning the parameters.

When there is no replication of observations and the errors are suitably restricted, there exists a class of consistent estimators which are ratios of linear forms in the observations. When partial information is available concerning the spacing of the X_i there are rational grounds for choosing a suitable member of the class.

BIBLIOGRAPHY

[1] Bartlett, M. S. [1949]. Fitting a straight line when both variables are subject to error. *Biometrics* 5.

[2] Dorff, M. R. [1960]. Large and small sample properties of estimators for a linear functional relation. *Unpublished Ph. D. Thesis*, Iowa State University of Science and Technology.

[3] Dorff, M. R. and Gurland, J. [1961]. Estimation of the parameters of a linear functional relation, Accepted for publication in Journal of the Royal Statistical Society, Series B.

[4] Gibson, W. M. and Jowett, G. H. [1957]. 'Three-Group' regression analysis. Part I. Simple Regression Analysis. *Appl. Statist. 6.*

[5] Gurland, J. [1956]. *Paper Presented at meeting of Institute of Mathematical Statistics Chicago.*

[6] Hooper, J. W. and Theil, H. [1958]. The extension of Wald's method of fitting straight lines to multiple regression. *Rev. Int. Statist. Inst. 26.*

[7] Housner, G. W. and Brennan, J. F. [1948]. Estimation of linear trends. *Ann. Math. Statist. 19.*

[8] Nair, K. R. and Shrivastava, M. P. [1942]. On a simple method of curve fitting. *Sankhya 6.*

[9] Theil, H. and van Yzeren, J. [1956]. On the efficiency of Wald's method of fitting straight lines. *Rev. Int. Statist. Inst. 24.*

[10] Wald, A. [1940]. The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist. 11.*

STATISTICS FOR A DIAGNOSTIC MODEL[1]

ADRIANUS J. VAN WOERKOM AND KEEVE BRODMAN
Cornell University Medical Center
New York, New York, U. S. A.

INTRODUCTION

In recent years, several methods have been proposed for making medical diagnoses by machine (Ledley and Lusted [1959], Crumb and Rupe [1959]). A method devised by Brodman *et al.* [1959, 1960] has been used to program a high-speed electronic computer for making presumptive medical diagnoses using only information relating to the age, sex, and responses of patients to a standardized health questionnaire.

The method assigns patients to none, one, or several of 60 selected disease categories. The 60 diseases most frequently diagnosed by hospital physicians in men and the 60 diseases in women were chosen for study. The method was developed with data referring to 5,929 consecutive adult white patients (2,718 men and 3,211 women) admitted to the outpatient departments of The New York Hospital, a large general hospital, during the 18-month period beginning July 1, 1948. It was tested with data referring to 2,745 consecutive adult white patients (1,280 men and 1,465 women) admitted during the 12-month period beginning January 1, 1956.

Each patient's symptoms were elicited through a printed form, the Cornell Medical Index-Health Questionnaire (CMI). The CMI was devised to collect diagnostically important elements of the medical history given by general medical patients, without expenditure of a physician's time. Solely with these data, a physician can often correctly predict which diseases will be found in subsequent examination (Brodman *et al.* [1951]).

Additional data abstracted for analysis from the hospital records include each patient's sex and age, along with the diagnoses made by hospital physicians after eliciting a history and performing physical and laboratory examinations. These diagnoses for the 1948–1949 data were coded according to the U. S. Public Health Service *Manual for Coding Causes of Illness* [1944] and form the standard against which the method is evaluated.

[1] Leo S. Goldstein and Albert J. Erdmann Jr. participated in this study, which was carried out at the Department of Medicine, New York Hospital-Cornell Medical Center with the support of a Public Health Service Research Grant.

The problem may be phrased in the following terms: Patients in the sample may belong to a number of specified disease categories. Patients were assigned by physicians to disease categories with a low incidence of false positives (incorrect diagnoses) but were not assigned to every category to which they belonged. Each patient has a number of attributes (complaints on the CMI). These attributes may or may not be correlated with each other. No one-to-one relation exists between categories and attributes. The problem is to assign patients to categories solely by means of attributes indicated on the CMI, with a maximum of correct and a minimum of incorrect assignments.

There are two obvious approaches to the solution of this problem:

1. The use of the conditional probabilities of a patient's having an attribute, given the category to which he belongs. From these, using Bayes' theorem, the conditional probability can be found for the patient's belonging to a specific category, given the attributes (Feller [1957], Chapter 5).

2. The use of significance values or weights for the attributes with respect to the categories.

CONDITIONAL PROBABILITIES

Let P_k be the probability or relative frequency of attribute A_k in the total sample, and p_{jk} the conditional probability of the attribute A_k given the category D_j. Then

$$p_{jk} = \Pr(A_k \mid D_j). \tag{1}$$

If n_j is the sub-sample size of category D_j and N the size of the total sample, then

$$\Pr(D_j) = n_j/N. \tag{2}$$

The conditional probability for a group of attributes $A_{k_1}, \cdots, A_{k_n}$ is:

$$\Pr(A_{k_1} \cdots A_{k_n} \mid D_j) = \Pr(A_{k_1} \mid D_j) \Pr(A_{k_2} \mid D_j) \cdots \Pr(A_{k_n} \mid D_j). \tag{3}$$

Using Bayes' theorem, the probability of the patient's belonging to category D_j given the attribute complex $A_{k_1}, \cdots, A_{k_n}$ is:

$$\Pr(D_j \mid A_{k_1} \cdots A_{k_n}) = \frac{\Pr(A_{k_1} \mid D_j) \cdots \Pr(A_{k_n} \mid D_j) \Pr(D_j)}{\sum_i \Pr(A_{k_1} \mid D_i) \cdots \Pr(A_{k_n} \mid D_i) \Pr(D_i)}. \tag{4}$$

The expression (4) for Bayes' rule will have to be adjusted for correlation between the attributes. As can be seen from Table 1, the sample sizes of the 60 selected diseases are generally quite small. These sample sizes are of the order of 30, which makes the measurements of the

TABLE 1

Sample Sizes of 60 Diseases in Men and Women.

Sample	Men	Women
Size	*N*	*N*
10–19	16	8
20–29	13	12
30–39	8	14
40–49	6	6
50–99	10	12
100–149	6	5
150–404	1	3

correlation coefficients uncertain. In Table 2 a tabulation is given for the correlation coefficients for the symptoms with higher proportionality in the disease sub-samples than in the total sample. The coefficients are reduced to unit variance. Only 22 percent of these correlations are more than twice their standard error.

In general, a patient has more than two symptoms and multiple correlations will have to be considered. Because of the small sub-sample sizes, these multiple correlations, either directly observed or

TABLE 2

Distribution of Item Inter-Correlations Reduced to Unit Variance.

Correlation/Variance	*N*
4.0-6.0	12
3.0-4.0	20
2.0-3.0	43
1.5 2.0	43
1.0-1.5	41
0.5-1.0	57
0.0-0.5	38
0.0-(−0.5)	44
(−0.5)-(−1.0)	24
(−1.0)-(−1.5)	13
(−1.5) (−2.0)	3
(−2.0) (−3.0)	0
(−3.0)-(−4.0)	0
(−4.0)-(−6.0)	0

TABLE 3

Number of Diseases Diagnosed Per Patient (Maximum of 3).

Number of Diseases	*N*		%	
	Men	Women	Men	Women
1	1456	1635	54	51
2	646	811	24	25
3	616	765	22	24

derived from the covariances, will be more uncertain than the zero order correlations and their use could lead to a large amount of bias. In view of this uncertainty and the tremendous amount of work involved, the measurement of all the correlations needed for the proper application of Bayes' rule is not justified.

Of particular interest with respect to the use of Bayes' rule, many patients have more than one disease. In Table 3 a distribution of diagnosed diseases is given for the 1948–1949 New York Hospital population.

The presence in a patient of more than one disorder considerably complicates a method using conditional probabilities. Symptoms claimed by a patient with a particular diagnosed disease may be related to other and often undiagnosed diseases. Due to a random distribution of most diseases with respect to each other, the co-existence of diseases in a patient will not seriously obscure the cluster of symptoms having substantially higher proportionality within a disease category than in the total sample, nor will the co-existence of diseases seriously distort the correlations of these symptoms within the disease category. Correlations between symptoms which have an incidence within the disease category similar to the incidence in the total sample, however, may be due to other diseases.

Some of these difficulties could be avoided by studying disease complexes rather than single diseases. It is evident from the sample sizes of single diseases as shown in Table 1, however, that the study of disease complexes in the sample available would not have been possible because of the small sub-sample sizes for these complexes. Even in a sample much larger than the one available, compounding of error because of false negatives would still be uncontrolled.

The effect on the conditional probabilities of the presence of diseases other than the one being studied would be partially eliminated by using only those attributes which have a substantially higher incidence within

TABLE 4

Hypothetical Sample for Conditional Probabilities.

	D_1	D_2
n	90	10
A_1	33%	60%
A_2	33%	60%
$n(A_1A_2)$	9	4
$\Pr(D\|A_1A_2)$	9/13	4/13

the disease category than in the total sample. This approach would also eliminate those symptoms which occur less frequently in the disease than in the total sample. (In clinical practice, these symptoms are given little diagnostic weight.)

Another objection to the use of Bayes' rule is that the conditional probabilities for a disease, given a set of attributes, is directly proportional to the sub-sample size of the disease category. (The medical profession first considers diseases with a high incidence but does not make a diagnosis on this basis.)

If the data available are consistent and complete, Bayes' rule will give a mininum number of wrong assignments. If the data are inconsistent and incomplete, as in the sample studied, there may be many cases of wrong assignments and methods other than Bayes' might be more effective. Referring to Table 4, if the 9 cases in D_1 with attributes A_1 and A_2 were false negatives in the data with regard to D_2 , the hypothesis that all patients having A_1 and A_2 belong to D_2 will be more effective than the use of Bayes' rule. This hypothesis can be tested only by re-examining hospital records or re-calling patients.

Notwithstanding the objections raised, the conditional probability method, with the modifications described above, was tested with the 1948–1949 sample of male patients. The results showed a number of expected inconsistencies. Nearly always, those categories with the largest number of attributes with high relative frequencies were favored. In addition, categories with small sample sizes and with a small number of attributes of high relative frequency (that would have been identified by a physician in many cases) failed to get probabilities high enough to be considered.

The most reliable information in the data is the diagnoses made by physicians after they have examined the patients. The one thing that can be established with minimum error from the data studied here is the significance of an attribute with respect to a disease. The conditional probability approach fails to make use of this information.

It was found from clinical experience and from noting the relative frequencies or proportionalities P_k and p_{ik} that a disease category is characterized by a cluster of attributes each with a much higher proportionality in the disease category than in the total sample. If a patient's cluster of complaints is similar to the cluster of complaints of the average patient in that disease category, he can be assigned to the category. This approach is discussed in the following section.

SIGNIFICANCE VALUES

Let the variable S'_{jk} be given as

$$S'_{jk} = n_j(p_{jk} - P_k)/\sqrt{n_jP_k(1 - P_k)}, \qquad (5)$$

where the notations are as before. It can be shown that in a sample where the attribute A_k is randomly distributed with probability P_k, the variable S'_{jk} is approximately normally distributed if $n_jP_k(1 - P_k)$ is not too small. However, according to Feller [1957 p. 170], the error committed in replacing the binomial distribution by the normal is surprisingly small, even for values of $n_jP_k(1 - P_k)$ of the order of 1.5. The comparison between the two for the probabilities of the variable between two boundaries is by far not as good, but this is of little consequence in this case. Then in the sub-sample of size n_j, the probability of observing the value S'_{jk} by chance is proportional to the expression $\exp(-S'^2_{jk})$. The observed value of S'_{jk} in the sub-sample of category D_j therefore expresses a measure of significance in the usual sense. Because the constant n_j occurs in the significance values for all attributes with regard to the category D_j, this factor can be dropped from the expression. Then,

$$S''_{jk} = (p_{jk} - P_k)/\sqrt{P_k(1 - P_k)} \sim (p_{jk} - P_k)/\sqrt{P_k}\,. \qquad (6)$$

The term $(1 - P_k)$ varies between .99 and .46 with nearly all values around .80 with the square root equal to .9. By ignoring the term $(1 - P_k)$ an error is made of approximately 10 percent. This error is of little consequence because of statistical inaccuracies in the other parameters.

Let all attributes A_k for which $S''_{jk} \geq 2$ be called significant attributes. If the complex of significant attributes contains more than one attribute, $\sum S''^2_{jk}$ represents a measure of the joint probability of the attribute complex and thus the significance of the complex. The significance value of the complex should also include the correlation between the attributes. As mentioned before, however, in the sample under study these correlations are generally insignificant and a linear model can be used.

The variable S''_{jk} measures not only the significance of the attributes but also their correlation with respect to the categories. The observed coincidence of category D_j and attribute A_k is equal to $(n_j/N)p_{jk}$ and the product of the expected values is $(n_j/N)P_k$. The co-variance is then equal to

$$(n_j/N)p_{jk} - (n_j/N)P_k = n_j/N(p_{jk} - P_k). \tag{7}$$

In order to reduce (7) to the correlation coefficient, the expression has to be divided by the square root of the product of the variances of $n_j \ N$ and P_k. The factor n_j/N is constant for all attributes in the same category and therefore can be eliminated. The variance of P_k depends on the universe from which our sample is drawn. If the universe has a Poisson distribution the value is equal to P_k; if the universe is binomial the value is equal to $P_k(1 - P_k)$.

The resulting expressions

$$(p_{jk} - P_k)/\sqrt{P_k} \quad \text{or} \quad (p_{jk} - P_k)/\sqrt{P_k(1 - P_k)}$$

are identical to S''_{jk} (6).

The following form was adopted for S_{jk}:

$$S_{jk} = [(p_{jk} - P_k)/2\sqrt{P_k}] - 1, \tag{8}$$

where the factor 2 in the denominator yields a convenient scale. A minimum value of 3 was originally set for $\sqrt{P_k}$. Further study of the data indicates that setting a minimum value for this is unnecessary, and that all negative values can be set to zero.

If the attributes $A_{k_1}, \cdots, A_{k_n}$ are present in a patient, the corresponding sums

$$(Sc)_j = \sum_{k_1}^{k_n} S_{jk}^2 \qquad (j = 1, \cdots, 60) \tag{9}$$

are called the scores. The scores for a patient are normalized by dividing them by the mean score for the entire category. This normalization compares the patient to the average patient. If the sample sizes had been kept in the variables S_{jk} they would have cancelled out at this point.

It cannot be assumed that the score as indicated in formula (9) will give optimal results. The score is essentially a measure of the probability of the chance existence of any sub-sample with the same relative frequency p_{jk} for the attributes $A_{k_1}, \cdots, A_{k_n}$ as in the subsample of category D_j. When the score is high this probability is low, that is, a high score indicates a small probability for the patient's not belonging to the category. Had the linear sum of the significance values S_{ik} been used to determine the score, small changes in the significance values would be reflected by large changes in the probabilities. This would make the score a bad indicator of the probabilities. If powers of S_{ik} higher than the second were used, the score would be a bad indicator for the opposite reason.

Various models (linear sum, sum of squares, sum of higher powers, etc.) may be compared quantitatively only if the distribution of scores in both the restricted and the category sub-samples are known. By restricted sample is meant the total sample minus the category sub-sample. These distributions could not be obtained for the sample studied, because the restricted sample contains patients who really belong to the category sub-sample but have not been assigned to it in the hospital (false negatives). Some of the disease categories, further, are significantly correlated. These correlated category sub-samples must be removed from the restricted sample in order to get the desired distribution. Removal of false negatives and adjunct category sub-samples is impossible for a large sample.

Physicians, when making a diagnosis, consider the patient's age. Table 5 shows the age distribution by decades of the 15 diseases diagnosed most frequently in men and in women in the 1948–1949 population. The relative frequency of a disease category is a function of age. The use of these relative frequencies in the sense of relative probabilities is not consistent with a model based on significance values of the attributes. There is, however, another aspect to be considered. Let the sample be subdivided into a number of age groups. The relative sub-sample sizes $(r_i)_m$ for disease category D_i and age group m will in general not be constant with respect to m. The significance value of an attribute with respect to a category is a function of the sub-sample size. If no subdivision of the total sample is considered, the effect of the sub-sample size is eliminated because the score of the patient is always compared to the score of the average patient and the sub-sample size does not have to enter into the significance value. If the same comparison is made in case a subdivision exists, the sub-sample size will have to be considered in the significance value. The expression for $(S_{ik})_m$ will be

$$(S_{ik})_m = \frac{\sqrt{(r_i)_m}(p_{ik} - P_k)}{\sqrt{P_k(1 - P_k)}} \sim \frac{\sqrt{(r_i)_m}(p_{ik} - P_k)}{\sqrt{P_k}}. \tag{10}$$

The mean scores for each age group within the disease category were computed to test the possibility that the relative frequency p_{ik}

of the attributes within the category sub-sample is a function of age. No significant interaction was found between the mean scores and age.

In order to smooth the relative frequencies of the disease categories with respect to age, moving averages of three successive decades were used. After the $\sum S_{ik}^2$ scores were formed, using formula (8) for S_{ik}, the scores were multiplied by $(r_i)_m$.

This modification of the significance value changes the distributions of the scores for both the categories and total sample so that more success can be expected from the model. Indications of this are made in Figure 1, where *I* is the distribution curve for the restricted sample and *II* is the distribution curve for the sub-sample category. The dotted lines show the distribution curves using the formula modified for age. The trend will be primarily towards a smaller number of wrong assignments (false positives).

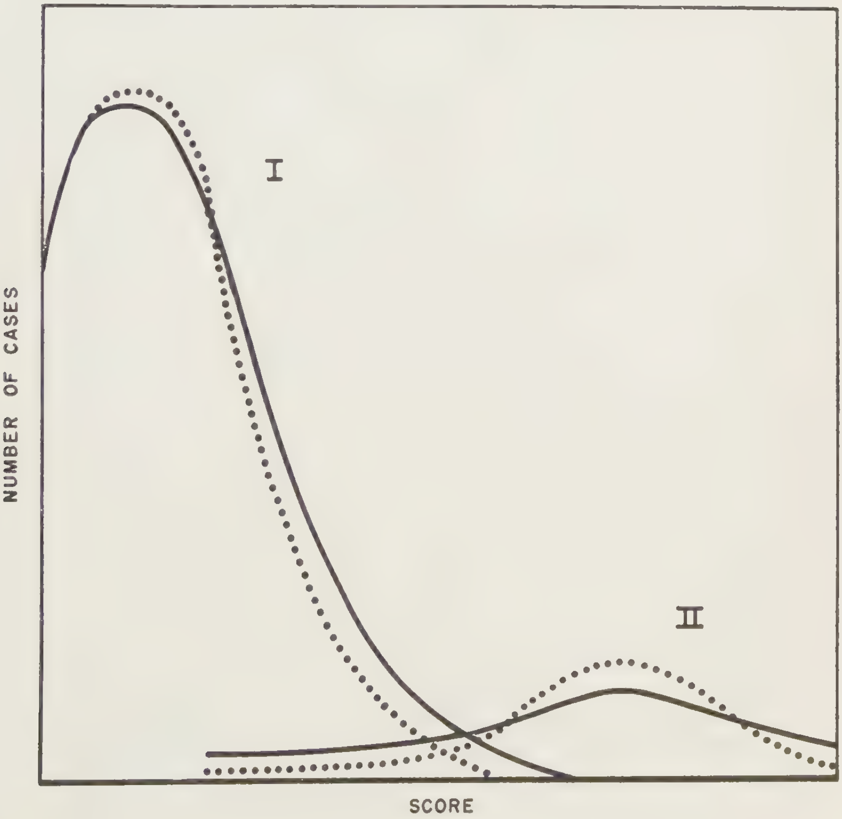


FIGURE 1

Distribution of Scores for Category and Total Sample.

TABLE 5

Age Distribution of Patients With Selected Diagnoses.

MEN

			Percent of *N* by Decade						
Diagnosis	Code	*N*	10–9	20–9	30–9	40–9	50–9	60–9	70+
Total		2718	05	16	18	20	22	14	04
Psychoneurosis	330	233	04	26	24	26	15	04	00
Benign prostatic hypertrophy	630	140	00	00	02	07	34	37	20
Inguinal hernia	550	120	04	03	11	28	24	23	07
Coronary artery disease	382	110	00	00	01	19	25	37	18
Varicose veins	410	110	01	05	15	22	29	23	05
Duodenal ulcer	527	109	02	11	27	30	18	09	03
Sinusitis	495	100	12	21	21	19	22	04	01
Functional gastro-intestinal disorder	564	96	00	25	22	28	16	09	00
Hemorrhoids	415	83	04	11	24	23	25	12	01
Errors of refraction	343	80	15	26	16	25	10	06	01
Infected teeth	510	80	11	14	25	20	19	09	03
Osteo-arthritis	724	76	00	00	03	17	39	32	09
Hypertensive cardiovascular disease	370	72	00	00	07	21	40	26	06
Arteriosclerosis	400	65	00	00	00	08	29	46	17
Deflected nasal septum	496	62	10	32	24	18	15	02	00

TABLE 5 (*Continued*)

WOMEN

Diagnosis	Code	N	Percent of N by Decade						
			10–9	20–9	30–9	40–9	50–9	60–9	70+
Total		3211	07	19	25	23	11	12	02
Psychoneurosis	330	404	06	23	35	22	08	06	00
Obesity	241	202	08	13	20	26	16	15	00
Varicose veins	410	168	00	05	18	31	18	24	03
Functional gastro-intestinal disorder	564	112	02	19	35	24	13	06	02
Errors of refraction	343	111	17	23	14	25	10	12	00
Hypertensive cardiovascular disease	370	111	00	00	05	21	33	29	12
Cervicitis	652	110	02	22	38	34	05	00	00
Dermatitis venenata	718	103	09	15	19	20	19	13	05
Sinusitis	495	96	06	28	31	23	07	04	00
Osteo-arthritis	724	96	00	00	04	28	23	33	11
Hemorrhoids	415	74	00	18	30	32	07	11	03
Infected teeth	510	71	18	18	30	20	06	08	00
Biliary calculi	585	68	00	10	24	29	15	22	00
Menstrual disorders	664	67	16	33	25	25	00	00	00
Allergic rhinitis (hay fever)	500	64	17	44	20	13	03	02	02

Because the process of assignment by physicians at The New York Hospital for the sample of patients under study cannot have been entirely divorced from the attributes, bias in assignment may be expected with respect to the attributes. Bias in p_{ik} arises partly as a function of the clinical process used in making a decision. A physician in oral interview elicits a patient's symptoms and then investigates with physical and laboratory examinations the diseases he presumes to be associated with the symptoms. Some of these symptoms will be similar to those indicated by the patient on the CMI. The presumptive diagnoses suggested by the symptoms reported to the physician can thus be expected to be related to the diagnostic assignments made from the attributes on the CMI. Bias in p_{ik} arising from the clinical process is inherent in the data and cannot easily be corrected.

In Table 6 is given a sample of the relative frequencies for 27 selected items on the CMI with respect to 6 selected diseases for men in the 1948–1949 population. The numbers in the heading of the table correspond to the item numbers on the questionnaire. The relative frequencies are expressed in percent. In the 7th line of the table the corresponding relative frequencies are given for the entire 1948–1949 sample of male patients. The last six lines of the table give the corresponding significance indices (S_{ik}). As can be seen, the complaints for the 6 selected diseases have very little overlap and form a good basis for differentiation of the diseases. The disease categories and items in the table were selected to illustrate this characteristic. The table shows that differential classifications can be made on the basis of the significance indices.

The sample in Table 7 was chosen from the male 1948–1949 population to present a case where overlap among diseases exists in the complaints. In the first column of the table the item content is given with the item number from the questionnaire. Column headings identify the diseases. The last column gives the number of diseases for which the particular item is significant and the last line in the table shows the number of these items significant for the disease. All significance indices smaller than 3 are omitted from the table. Even with the overlap present, the table still shows that differentiation among diseases can be made on the basis of the significance indices. Table 8 gives a comparable illustration for the women in the 1948–1949 population.

Some points in Tables 7 and 8 are of interest:

1. In Table 7, item 56 of disease 121 ($N = 12$) illustrates bias introduced because of small sample size in the disease.
2. In Table 8, item 41 of disease 382 is related to age.

TABLE 6

PERCENT "YES" AND SIGNIFICANCE INDICES FOR ITEMS WITH LITTLE OVERLAP AMONG DISEASES

Disease	Item Number																										
	7	8	9	17	19	22	25	30	31	32	34	39	48	51	52	53	54	56	65	66	67	70	102	103	104	105	106
	Percent "Yes"																										
350	59	56	56	22	26	22	11	26	04	15	33	15	22	11	15	11	15	19	19	26	00	30	06	23	10	03	00
471	09	07	09	58	49	60	47	56	37	23	53	02	33	15	15	09	03	00	28	21	09	19	02	33	26	09	09
360	03	06	12	33	24	12	12	73	64	70	61	88	30	24	36	12	27	06	18	24	09	24	09	18	15	09	03
527	09	05	06	18	15	10	11	40	13	15	25	12	65	63	61	57	74	64	06	10	03	25	09	24	20	07	06
724	17	04	08	25	25	14	16	29	14	12	26	14	21	14	21	11	13	14	36	54	30	47	07	43	43	08	07
630	17	09	13	18	13	11	10	23	21	14	24	14	24	19	24	14	22	14	13	22	08	22	49	69	61	31	25
Total Sample	13	09	09	18	17	13	10	26	17	14	25	13	27	19	24	13	19	10	13	20	06	19	11	29	29	08	06
	Significance Index																										
350	6	7	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
471	0	0	0	4	3	6	5	2	2	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
360	0	0	0	1	0	0	0	4	5	7	3	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
527	0	0	0	0	0	0	0	0	0	0	0	0	3	4	3	5	5	8	0	0	0	0	0	0	0	0	0
724	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	3	2	0	0	1	0	0
630	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	2	3	2

TABLE 7

Significance Indices for Items With Overlap Among Diseases in Men

		Disease																Diseases
Item	Item	020	121	184	360	370	382	388	440	460	471	495	496	500	501	504	995	Per Item
Number	Content	Significance Index																
11	Lump in throat									3								1
12	Sneezing													5	3			2
13	Stuffed nose			5								3	3					3
14	Running nose													4	3			2
16	Frequent colds								3									1
17	Chest colds	3	4								4				4			4
19	Severe colds										3							1
20	Hay fever history													8	4			2
21	Asthma history													3	8			2
22	Constant coughing	5	4								6				5	3		5
23	Coughing blood	3																1
25	Chronic chest condition	4	3								5				4		3	5
26	Tuberculosis history	5															6	2
28	High blood pressure history					5												1

TABLE 7 (*Continued*)

30	Chest pains				4		4											2
31	Thumping heart				5		3	3										3
32	Rapid heart				7			3										2
33	Breathing difficulty						3								6	3		3
34	Shortness of breath				3										3			2
35	Shortness of breath at rest														5	3		2
39	Heart trouble history				8		5	3										3
45	Poor appetite		4															1
56	Stomach ulcer history		3															1
111	Fatiguability		3															1
125	Rheumatic fever history				5													1
132	Chronic disease history		3												3			2
	Items Per Disease	5	7	1	6	1	4	3	1	1	4	1	1	4	11	3	2	

TABLE 8

Significance Indices for Items With Overlap Among Diseases in Women

		Disease										Diseases
Item	Item	360	370	382	388	440	471	495	500	501	995	Per Item
Number	Content	Significance Index										
12	Sneezing								7	5		2
13	Stuffed nose					3	3	3	5			4
14	Running nose						3	4	5			3
16	Frequent colds						4			3		2
17	Chest colds						8			5		2
18	In bed with colds	3				3	3					3
19	Severe colds					3	6			3		3
20	Hay fever history								7	5		2
21	Asthma history								3	8		2
22	Constant coughing						8			7		2
25	Chronic chest condition						8			3		2
26	Tuberculosis history										5	1
28	High blood pressure history		5	4								2

TABLE 8 (*Continued*)

30	Chest pains	4												1
31	Thumping heart	4												1
32	Rapid heart	5												1
33	Breathing difficulty	3										4		2
34	Shortness of breath	3												1
35	Shortness of breath at rest	4												1
39	Heart trouble history	8		3	3									3
41	Missing teeth			3										1
84	Faint feeling				3									1
92	Nail biting				3									1
118	Sickly person												3	1
125	Rheumatic fever history	8												1
132	Chronic disease history	3												1
133	Underweight												3	1
137	Injury history				3									1
181	Emotional instability				3									1
	Items Per Disease	10	1	3	5	3		8	2		5	9	3	

TABLE 9

Scores for 6 Patients.

Sex—	M M M F F F		
Age—	66 35 26 22 32 32		
Number of Complaints—	68 11 85 52 47 27		
Disease	Score	Disease	Score
1	68 00 00 00 00 00	31	25 00 00 00 00 37
2	12 00 12 00 00 00	32	12 00 12 00 12 12
3	00 00 00 00 00 00	33	75 00 00 00 00 00
4	50 00 00 00 00 00	34	95 00 00 00 18 00
5	87 00 00 00 00 00	35	00 00 00 00 00 00
6	43 00 00 00 00 00	36	00 00 18 00 37 00
7	00 00 00 00 00 00	37	75 00 00 00 31 00
8	00 00 00 43 12 12	38	12 00 18 00 18 00
9	18 00 00 00 00 00	39	00 00 50 18 12 12
10	18 00 25 00 00 00	40	12 00 25 00 00 00
11	18 00 95 00 00 00	41	31 00 00 00 00 00
12	00 00 50 37 25 12	42	18 00 12 00 00 00
13	00 00 50 00 00 00	43	68 00 00 00 00 00
14	00 00 00 00 00 00	44	50 00 00 00 00 00
15	00 00 95 00 00 00	45	75 00 00 00 12 00
16	00 00 62 00 00 00	46	43 00 12 81 62 00
17	00 00 00 00 00 00	47	31 00 00 00 00 00
18	00 00 00 56 68 43	48	00 00 00 00 00 00
19	37 68 75 00 00 00	49	00 00 00 00 00 00
20	87 00 00 00 00 00	50	12 00 00 00 00 00
21	93 00 00 43 25 12	51	31 00 00 00 00 00
22	75 62 62 00 00 00	52	56 00 00 00 00 00
23	75 00 00 00 00 00	53	81 00 00 00 00 00
24	25 00 00 00 62 00	54	25 00 37 00 00 00
25	00 62 00 00 00 00	55	56 00 00 00 00 00
26	00 00 00 00 56 25	56	00 00 00 00 00 00
27	00 00 37 00 00 00	57	00 00 00 75 87 00
28	12 00 43 00 50 00	58	00 00 31 00 00 00
29	87 00 12 00 31 95	59	18 00 00 00 00 00
30	37 00 50 00 00 00	60	00 00 00 18 31 00

3. In Table 8, item 28 of disease 382 represents the co-existence of disease 370.

In the study of Brodman *et al.* [1959], the $\sum (S_{ik})^2_m$ model was used to diagnose diseases in the 1956 sample of patients. A constant critical score of 35 for all categories was adopted, after examination of the scores obtained for the 1948–1949 sample. The use of critical scores determined for each disease independently may improve the model.

Table 9 gives the scores for 6 randomly selected patients diagnosed in the hospital as having rheumatic heart disease. The heading of the table gives the sex and age of the patients and the number of "yes" responses to the questionnaire. The number of diseases indicated for the 6 cases in the table ranges from 21 for the first case to 3 for the second and for the last case.

The model developed with the S_{ik} values derived from the 1948–1949 sample of patients was used to program a data processing machine (van Woerkom [1960]) and was tested with the 1956 sample of patients. The machine made correct assignments in 44 percent of cases where the 60 diseases were diagnosed in the hospital. In order to test for false positives, that is, diseases diagnosed by the machine when the disease was not present, a sample of 350 patients was selected for which the hospital records could give evidence for the presence or absence of a disease. In only 5 percent of cases did the model generate false positives. Results with the 350 patients are essentially the same as those obtained by a physician experienced in the interpretation of the CMI, except that for the diagnosis of psychoneurosis the physician made correct assignments more often than did the machine (81 percent as compared to 42 percent).

CONCLUSIONS

An approach used in efforts such as the one presented in this paper is best guided by the characteristics of the data interpreted. It is primarily because of the many false negatives in the large amount of data analyzed that the present model was developed. When small amounts of accurately defined data are analyzed the method may require revision. This paper presents a justification of the approach reported here, on an analytic basis and on the basis of the results obtained showing few false positives among many correct assignments.

BIBLIOGRAPHY

Brodman, K., Erdmann, A. J., Jr., Lorge, I., and Wolff, H. G. [1951]. The Cornell Medical Index—Health Questionnaire II. As a diagnostic instrument. *J.A.M.A* *145*, 152–7.

Brodman, K., van Woerkom, A. J., Erdmann, A. J., Jr., and Goldstein, L. S. [1959]. Interpretation of symptoms with a data processing machine. *A.M.A. Archives of Internal Medicine. 103*, 776–82.

Brodman, K. [1960]. Diagnostic decisions by machine. I.R.E. *Trans. Medical Electronics. ME-7*, 216–9.

Crumb, C. B., Jr., and Rupe, C. E. [1959]. The automatic digital computer as an aid in medical diagnoses. *Proceedings of the Eastern Joint Computer Conference*, 174–9.

Feller, W. [1957]. *Probability Theory and Its Applications*, 2nd. edition. New York, John Wiley and sons.

Ledley, R. S., and Lusted, L. B. [1959]. Reasoning foundations of medical diagnoses. *Science 130*, 9–21.

van Woerkom, A. J. [1960]. Program for a diagnostic model. *I.R.E. Trans. Medical Electronics. ME-7*, 220.

QUERIES AND NOTES

D. J. Finney, *Editor*

159 NOTE: Three-Quarter Replicates of 2^4 and 2^5 Designs

Peter W. M. John

California Research Corporation, Richmond, California and University of California, Berkeley, California. U. S. A.

1. THE 2^4 DESIGN

The usual half replicate of the 2^4 design is defined by $1 = \pm x_1x_2x_3x_4$; the main effects are clear of two-factor interactions, but the two-factor interactions themselves are confounded in pairs. The three-quarter replicates given in this section enable both the main effects and two-factor interactions to be estimated clear of two-factor interactions.

The designs are obtained by omitting any one of the following quarter replicates, defined by $x_4 = \pm x_2$, $x_3 = \pm x_1x_2$.

I. 1, *ac*, *bcd*, *abd*; $x_4 = x_2$, $x_3 = -x_1x_2$.
II. *a*, *c*, *abcd*, *bd*; $x_4 = x_2$, $x_3 = x_1x_2$.
III. *b*, *abc*, *cd*, *ad*; $x_4 = -x_2$, $x_3 = x_1x_2$.
IV. *ab*, *bc*, *acd*, *d*; $x_4 = -x_2$, $x_3 = -x_1x_2$.

We consider the case in which I is omitted.

The three quarters may be combined into three half replicates; II, III, $x_3 = x_1x_2$; II, IV, $x_4 = x_1x_3$; III, IV, $x_4 = -x_2$.

We estimate A from the half replicate III, IV, where $x_1 = -x_1x_2x_4$, so that *A* is confounded with *ABD*. In the other half replicates *A* is confounded with the two-factor interactions *BC* or *CD*. The other effects are obtained, each from a half replicate, in the same way and the analysis is presented in tabular form below. It may be shown that these are the least squares estimates, when higher-order interactions are assumed to be negligible. *BD*, and the mean, may be estimated from two half replicates; the least squares estimates are then the averages of these two estimates.

2. THE 2^5 DESIGN

The main effects and two-factor interactions in the 2^5 experiment can all be estimated from a half replicate. The variance of each esti-

Effect	Sets Used	Confounding
A	III, IV; $x_4 = -x_2$	ABD
B	II, IV; $x_4 = x_1x_3$	$ABCD$
C	III, IV; $x_4 = -x_2$	BCD
D	II, III; $x_3 = x_1x_2$	$ABCD$
AB	II, IV; $x_4 = x_1x_3$	BCD
AC	III, IV; $x_4 = -x_2$	$ABCD$
AD	II, III; $x_3 = x_1x_2$	BCD
BC	II, IV; $x_4 = x_1x_3$	ABD
BD	II, IV; $x_4 = x_1x_3$	ABC
	or II, III; $x_3 = x_1x_2$	ACD
CD	II, III; $x_3 = x_1x_2$	ABD

mate is $\sigma^2/4$, and all the degrees of freedom are used for effects. We present in this section a three-quarter replicate in which all the effects are estimable with variance $3\sigma^2/16$ and which provides 8 degrees of freedom for error.

The three-quarter replicate using three quarters defined by $x_5 = \pm x_2x_3$, $x_4 = \pm x_1x_2x_3$ is discussed in Davies [1956]. In this design all the main effects and two-factor interactions, except E, AD, BC, are estimated by averaging the estimates from two half replicates with variance $3\sigma^2/16$. The latter three effects, however, are estimated from only a single half replicate with variance $\sigma^2/4$.

If the quarter omitted is defined by $x_4 = \pm x_1$, $x_5 = \pm x_2x_3$, we have a design in which each main effect and each two-factor interaction is estimated from two half replicates. The particular design obtained by omitting the quarter $x_4 = -x_1$, $x_5 = +x_2x_3$ has the additional advantage that it may be developed by augmenting a "one-at-a-time" experiment, Daniel [1958].

3. SOME ADDITIONAL DESIGNS

The above designs possess a certain symmetry inasmuch as each factor appears at its high level and at its low level the same number of times. Another set of designs may be obtained when a half replicate, in which the highest-order interaction is confounded with the mean, is augmented by a quarter which also confounds a main effect. For example, one might take three of the sets of treatment combinations defined by

$$1 = \pm x_1 = \pm x_2x_3x_4(x_5) = \pm x_1x_2x_3x_4(x_5).$$

In this design the first factor appears at one level in 2^{n-1} points and at the other level in 2^{n-2} points.

It may nevertheless be shown that the method of analysis given in the previous sections still gives the least squares estimates. Such a design has the advantage in the 2^4 case over the design given in Section 1 that the effect which is estimated with greatest precision is a main effect rather than a two-factor interaction.

REFERENCES

[1] Daniel, C., [1958]. On varying one factor at a time. *Biometrics 14*, 430–1.

[2] Davies, O. L., [1956]. *The Design and Analysis of Industrial Experiments.* Oliver and Boyd, London, 472–5.

160 NOTE: On the Extension of Stevens' Tables for Asymptotic Regression

S. Lipton

University of New South Wales
Kensington, Sydney, Australia.

Concerning H. Linhart's note ([1960]. *Biometrics 16*, 125), I think it is worthwhile pointing out my own extension to Stevens' tables. This work has been reported by H. D. Patterson ([1956]. *Biometrics 12*, 389 and [1958]. *Biometrika 45*, 323) and was carried out at Rothamsted Experimental Station. The extended tables give, in Stevens' notation, the six F's for $r = .10(.01).90$ for each integral value of n from 3 to 12 inclusive. These values do not in the main overlap Linhart's work.

The tables proved surprisingly "popular" and I received several requests for copies. However, as the entries corresponding to large r for small n seem to be somewhat inaccurate, (apparently they are correct only to about 4 significant figures compared to more than 6 figures elsewhere) I am now recomputing the values. I intend to take the opportunity of further extending the range of n up to 15.

161 NOTE: Corrected Error Rates for Duncan's New Multiple Range Test

H. Leon Harter

Aeronautical Research Laboratory
Wright-Patterson Air Force Base, Ohio, U. S. A.

The error rates published by the author [3] for Duncan's new multiple range test were based upon critical values tabulated by Beyer [1] and by Duncan [2], some of which were in error, as reported earlier by the author [4]. On the basis of the corrected critical values for this

CORRECTED TABLE 1C

Minimum Values of α_L Corresponding to $\alpha_S = 0.05, 0.01$

$(\alpha_L)_{\min} \mid \alpha_S = 0.05$

m	$N = 2$	$N = 3$	$N = 4$	$N = 6$	$N = 10$	$N = 16$	$N = 25$	$N \to \infty$
2	.0500	.0500	.0500	.0500	.0500	.0500	.0500	.0500
3	.0496	.0443	.0425	.0411	.0402	.0396	.0395	.0391
4	.0463	.0395	.0372	.0354	.0342	.0338	.0334	.0329
5	.0429	.0357	.0334	.0315	.0301	.0298	.0295	.0290
6	.0400	.0329	.0305	.0287	.0275	.0269	.0266	.0261
7	.0375	.0305	.0283	.0265	.0254	.0248	.0245	.0239
8	.0354	.0286	.0265	.0247	.0236	.0231	.0228	.0223
9	.0335	.0270	.0250	.0233	.0222	.0216	.0214	.0209
10	.0320	.0256	.0237	.0221	.0210	.0206	.0203	.0199
12	.0294	.0236	.0216	.0202	.0192	.0188	.0185	.0181
14	.0274	.0219	.0202	.0188	.0179	.0175	.0172	.0168
16	.0257	.0206	.0189	.0176	.0168	.0164	.0162	.0158
18	.0244	.0196	.0180	.0167	.0160	.0155	.0153	.0149
20	.0233	.0186	.0172	.0160	.0153	.0148	.0146	.0142

$(\alpha_L)_{\min} \mid \alpha_S = 0.01$

m	$N = 2$	$N = 3$	$N = 4$	$N = 6$	$N = 10$	$N = 16$	$N = 25$	$N \to \infty$
2	.0100	.0100	.0100	.0100	.0100	.0100	.0100	.0100
3	.0098	.0085	.0081	.0077	.0075	.0074	.0074	.0073
4	.0089	.0072	.0067	.0063	.0061	.0060	.0059	.0058
5	.0079	.0063	.0058	.0054	.0052	.0050	.0050	.0049
6	.0071	.0056	.0051	.0048	.0045	.0044	.0044	.0043
7	.0065	.0050	.0046	.0043	.0041	.0039	.0039	.0038
8	.0059	.0045	.0041	.0039	.0037	.0036	.0035	.0035
9	.0054	.0042	.0038	.0036	.0034	.0033	.0033	.0032
10	.0050	.0039	.0035	.0033	.0031	.0030	.0030	.0029
12	.0044	.0034	.0031	.0029	.0028	.0027	.0027	.0026
14	.0039	.0031	.0028	.0026	.0025	.0024	.0024	.0023
16	.0036	.0028	.0025	.0024	.0023	.0022	.0022	.0021
18	.0033	.0026	.0024	.0022	.0021	.0020	.0020	.0020
20	.0030	.0024	.0022	.0020	.0019	.0019	.0019	.0018

CORRECTIONS TO TABLE 2

Comparison of Type II-III Error Rates for Single Classification

A. Values of $\beta_2' = (\beta_S')_{\max}$ for $\alpha = 0.05$

N	m	$\delta'=1.5$	$\delta'=2.0$	$\delta'=2.5$	$\delta'=3.0$
5	3	.464	.198	.0602	
	4	.467	.197	.0577*	.0129
	5	.472	.200	.0572	.0120
	6	.478	.203	.0574	
	12	.508		.0626	
	20	.532		.0685	.0123
7	4	.286	.0723	.0111	
		$\delta'=1.0$	$\delta'=1.5$	$\delta'=2.0$	$\delta'=2.5$
9	4	.535	.169	.0251	
	8	.577			
		$\delta'=0.5$	$\delta'=1.0$	$\delta'=1.5$	$\delta'=2.0$
13	3			.0496	
	4			.0537	
	5		.372	.0572	
	6		.384	.0604	
	8		.403	.0659	
16	4	.775	.257	.0214	.00046
25	3		.0774		
	4		.0858		
	5		.0927		
	6		.0986		
31	4	.573	.0385		
	5		.0422		

*For this case $\beta_3' = (\beta_X')_{\min} = .0818$

B. Values of $\beta_2' = (\beta_S')_{\max}$ for $\alpha = 0.01$

N	m	$\delta'=2.0$	$\delta'=2.5$	$\delta'=3.0$	$\delta'=3.5$
5	3	.508	.229	.0726	.0183
	4	.487	.211	.0630	.0143
	5	.479	.204	.0588	.0124
	6	.476	.201	.0568	.0114
	12		.208	.0567	.0101
	16	.498	.215	.0589	.0103
		$\delta'=1.5$	$\delta'=2.0$	$\delta'=2.5$	$\delta'=3.0$
7	3	.577	.234	.0556	.0088
9	3	.397	.0989	.0126	
	5	.411	.103	.0119	.00075
		$\delta'=1.0$	$\delta'=1.5$	$\delta'=2.0$	$\delta'=2.5$
11	8		.302	.0470	
	12	.762	.324	.0528	
	16	.776	.341	.0577	.0033
13	8	.667	.201	.0185	.00051
		$\delta'=0.5$	$\delta'=1.0$	$\delta'=1.5$	$\delta'=2.0$
16	5			.0897	.0035
	6	.932	.536	.0946	
	12	.947	.585	.116	
21	16	.929	.440	.0388	
41	3	.678	.0372		
	5		.0466		
	8	.750	.0565		
	12		.0668		
	16		.0743		
61	10	.588	.0057		

test recently published by the author [5], Table 1C of [3], which gives minimum values of the Type I error rate α_L for the *LSD* test corresponding to $\alpha_S = 0.05, 0.01$ for Duncan's test, has been completely recomputed, and the corrected table is given below. Also given is a list of corrections to Table 2 of [3], which gives values of $\beta_2' = (\beta_S')_{\max}$, the maximum combined Type II–III error rate for Duncan's test, for $\alpha = 0.05, 0.01$. To reduce the length of the list, corrections of a single unit in the last place have been omitted, except when there are larger corrections for other value (s) of δ' for the same values of N and m. [In both tables, N represents the size of each sample, and m represents the number of sample means being compared. In Table 2, δ' represents the ratio of the difference between the largest and smallest sample means in the group being tested to the pooled estimate s of the population standard deviation σ]. Table 3 of [3], which gives maximum required sample sizes $(N_S)_{\max}$ for Duncan's test for prescribed values of α, β, and m, has not been recomputed on the basis of the corrected critical values. If this were done, there might be an occasional change of 1, but most values are correct as they stand.

REFERENCES

[1] Beyer, William H. [1953]. Certain percentage points of the distribution of the studentized range of large samples. Virginia Polytechnic Institute Technical Report No. 4.

[2] Duncan, David B. [1955]. Multiple range and multiple *F* tests. *Biometrics 11*, 1–42.

[3] Harter, H. Leon [1957]. Error rates and sample sizes for range tests in multiple comparisons. *Biometrics. 13*, 511–36.

[4] Harter, H. Leon. [1957]. Critical values for Duncan's new multiple range test (abstract), *Jour. Amer. Stat. Assoc. 52*, 372.

[5] Harter, H. Leon. [1960]. Critical values for Duncan's new multiple range test. *Biometrics 16*, 671–85.

162 QUERY: Multiple Comparisons between Treatments and a Control

For multiple comparisons between p treatments and one control Dunnett [1955] has tabulated values of t from which confidence limits can be calculated. His tables provide for values of p from 1 to 9, and for degrees of freedom from 5 to infinity. The full tables allow for different numbers of tests (n) for each treatment, but when the number is the same for each treatment the number of degrees of freedom is given by $(p + 1)(n - 1)$, so that a simpler table can be derived for p against n, including the following:–

	$P = 0.05$				$P = 0.01$			
n	$p = 2$	4	7	9	$y = 2$	4	7	9
2		2.85	2.81	2.81		4.43	3.96	3.83
3	2.34	2.47	2.56	2.60	3.61	3.45	3.39	3.38
4	2.18	2.36	2.48	2.54	3.19	3.20	3.22	3.24

Does it really make sense for an increase in the number of treatments in each test to lead to smaller significant differences when there are only two tests of each, but larger when there are four or more? Or, when there are three tests of each, for an increase in treatments to increase the size of differences which are significant at $P = 0.05$ but reduce those at $P = 0.01$?

REFERENCE

Dunnett, C. W., [1955]. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Stat. Assoc. 50*, 1096.

ANSWER:

At first sight the behaviour of the table does appear rather odd. Intuitively one would probably expect an increase in the number of treatments to lead to a larger significant difference in all circumstances. This would certainly be the case if the only error variation were that associated with the control and treatment means. But it must be remembered that the variance has to be estimated, and the value of t has to allow for error in its value too. The reason the value of t decreases with p for the small values of n is that there the number of degrees of freedom for estimating the variance is small and, in effect, we gain more from the increase in degrees of freedom than we lose as a result of the greater number of treatment means.

If the table were extended to larger values of p, the value of t for $n = 2$, and for $n = 3$ when $P = .01$, would pass through a minimum and then increase with the value of p, indicating that a point is reached at which any further increase in the number of degrees of freedom is not sufficient to compensate for the increase in the number of treatments. For values of $n \geq 4 (P = .01)$ and $n \geq 3 (P = .05)$, this is the situation which pertains for *all* values of p.

It may be verified that the same phenomenon occurs in the case of the Studentized range, if the percentage points of that distribution are tabulated with n as parameter instead of the number of degrees of freedom.

C. W. Dunnett
Lederle Laboratories Division
American Cyanamid Company
Pearl River, New York

163 QUERY: Error Rates in Multiple Comparisons

Many biologists, realising the necessity for statistical examination of data, are unfortunately without access to professional advice, and must perforce search the literature themselves. One such worker encountered Steel's description [1959] of a multiple comparison sign test, which seemed to fill an urgent need, yet on closer study to reveal an apparent paradox. As his example Steel applies his test to data from 13 soybean yield trials of 3 varieties, and concludes that one variety (Z) was significantly better than control (C) because it was inferior ($-$) to control in only 2 of the trials. If, however, there had been five varieties in each trial, Z would not have been significantly different from control, because with 5 treatments the number of minus signs must not exceed one for significance. And yet the inclusion of two more varieties in the trials could not have made one iota of difference to the yield figures for Z and C from which alone the number of inferiorities was obtained. How can the presence or absence of data about W, Y, etc., entirely irrelevant to the $Z : C$ comparison and to all the computations involved alter the probability that Z is the same as C? Or are there implied assumptions, not apparent to the amateur reading Steel's paper, about the behaviour of W and Y relative to Z and C which determine when this test should be used in preference to Dixon and Mood's [1946] simple sign test?

REFERENCES

Dixon, W. J., and Mood, A. M., [1946]. The statistical sign test. *J. Amer. Stat. Assoc. 41*, 557.

Steel, R. G. D., [1959]. A multiple comparison sign test: treatment versus control. *J. Amer. Stat Assoc. 54*, 767.

ANSWER:

This query seems to arise from a failure to recognize that several definitions of error rate exist. These, of which only two are considered here, were introduced in connection with multiple comparison procedures and are just now finding their way into application and textbook. The decision as to which error rate is more appropriate and what significance level is to be used must, finally, be made by the experimenter.

Error rates are defined on the basis that the null hypothesis is true.

Per-comparison error rate =

$$\frac{\text{Number of comparisons falsely declared significant}}{\text{Total number of comparisons}}.$$

Here, the comparison is the conceptual unit. When comparisons are chosen without reference to the experimental results, a test such as the (parametric) *lsd* provides a per-comparison error rate at the tabulated value regardless of whether or not the comparisons are independent. A criticism of tests using this error rate is that, if many comparisons are made, the probability of finding at least one falsely significant difference is high.

Experimentwise error rate =

$$\frac{\begin{array}{c}\text{Number of experiments with at least one difference}\\ \text{falsely declared significant}\end{array}}{\text{Total number of experiments}}.$$

Here, the experiment is the conceptual unit. For pairwise comparisons, we agree to test the largest difference when it occurs. If this difference is significant, then this experiment is one for which at least one difference has been declared significant. Tukey's (parametric) test based on the studentized range has such an error rate; Dunnett's (parametric) test of treatments against control also has this error rate.

Clearly, tests with an experimentwise error rate will require a larger difference for significance than that required by the *lsd*. Furthermore, the required difference will increase with the number of treatments. This becomes a basis of criticism: the probability of detecting a real difference of fixed size is lower when an experimentwise error rate is used and decreases with the number of treatments.

Error rates are not, perhaps, as different as they first seem. If α_C and α_E represent per-comparison and experimentwise error rates respectively, then for k independent comparisons $(1 - \alpha_E) = (1 - \alpha_C)^k$. For dependent comparisons, the relation is not as obvious. Harter

[1957] has tabulated comparisons of significance levels for a number of multiple comparisons test procedures intended to compare all pairs of means.

It is apparent, then, that if an investigator chooses a definition of error rate, a test, and a significance level, then this information permits him to compute the significance level for any other suitable test and error rate. This knowledge eliminates any paradox.

No statistical assumptions determine which test is appropriate, regardless of whether the test is parametric or non-parametric. One chooses an error rate and significance level, prior to the conduct of the experiment, to give the kind of protection desired. It seems to me that the basic problem concerns a power function, presumably related to "urgent need" and, in turn, the sample size necessary to accomplish clearly stated aims.

REFERENCE

Harter, H. Leon, [1957]. Error rates and sample sizes for range tests in multiple comparisons. *Biometrics 13*, 511–36.

R. G. D. Steel
Institute of Statistics
North Carolina State College
Raleigh, N. C., U. S. A.

BOOK REVIEWS

J. G. Skellam, *Editor*

Members are Invited to Suggest Books for Listing or Review to the Editor

10 FALCONER, D. S. **Introduction to Quantitative Genetics.** Edinburgh and London: Oliver & Boyd Ltd., 1960. Pp. ix + 365. Figures and Tables. 35s.

W. F. Bodmer, *Department of Genetics, University of Cambridge, Cambridge, England.*

There has long been a need for a good introductory textbook on quantitative genetics, and Dr. Falconer's new book satisfies this need more than adequately. The term *quantitative genetics* is here used in its broadest sense to include both the study of genetic changes in populations and the study of the inheritance of continuous variation. The genetic analysis of changes in quantitative characters in both natural and artificial populations requires a close interplay between these two aspects.

The first five chapters of the book are concerned with the analysis of changes in gene frequency in natural and artificial populations. These provide perhaps the best elementary exposition I know of what is usually called *population genetics*. Dr. Falconer has a great gift for lucid and stimulating exposition. He presents the essential mathematics in a way that should be intelligible event to the least mathematically minded biologist. The numerical examples are excellently chosen to help in the understanding of the mathematical formulae as are also the graphical illustrations. In my opinion Dr. Falconer rightly assumes a knowledge of elementary genetics and later also of the elements of statistics. His book would have been unnecessarily burdened by any attempt to make it complete in these respects.

The genetics of continuous variation is treated in the following fifteen chapters with the same high standard of exposition. This subject has been approached in a variety of ways by different research schools and suffers from a confusion of terminology and notation. Dr. Falconer sets himself the difficult task of integrating the subject at an elementary level and is on the whole most successful. It is not surprising if the subject matter is somewhat biased toward the approach of the Edinburgh 'school', of which he is a member. Dr. Falconer is at least honest in delineating controversial topics and making it quite clear when a discussion is more a matter of personal opinion and preference than of established fact.

Any criticism of the subject matter is perhaps more a criticism of the development of the subject than of Dr. Falconer's account of it. However there is a notable lack of any emphasis on the importance of linkage, and the treatment of Mather's concept of *balance* is rather brief. The importance of epistacy is also, in my opinion, underestimated. A good epistatic genotype may be quite rare, and so contribute only little to an interaction variance, and yet be of considerable importance to the practical breeder. As Dr. Falconer says in his final discussion, there is need for much more experimental work. This should not, I feel, be directed at estimating more and more heritabilities and variance components for a wide selection of characters and organisms, as he suggests, but aimed at a more penetrating and specific analysis of the fundamental mechanisms underlying the inheritance of quantitative characters.

In conclusion, this is a book that should appeal to a wide audience. For the experienced geneticist it provides a stimulating and refreshing review of the subject, and for the newcomer to genetics it offers an excellent introduction. There is no doubt that it will be the best standard textbook of its kind for some years to come.

11 GOLDBERG, S. **Probability: An Introduction.** Englewood Cliffs, N. J.: Prentice Hall, Inc. 1960. Pp. xiv + 322. 30 Figures, 44 Tables. $5.95.

F. N. David, *Department of Statistics, University College, London, England.*

This book, written by a pupil of W. Feller, has been tried out, we are informed, on the freshman mathematics class of Oberlin College. This should give some idea of the level of mathematical sophistication which is aimed at and which the author generally achieves.

The first chapter on the manipulations and the algebra of sets is necessary for the main development of the book. It will be found difficult for persons wishing to learn probability with a view to practical applications, but the would-be mathematician will find it interesting, for the elements of mathematical logic are introduced. Probability in finite sample spaces follows. The author does not really attempt to justify the calculation of a probability, being concerned simply to define a one to one correspondence between the field of events and the probability set, and then to discuss the sub-sets involved. He says: "The theory of probability begins when a sample space, S, the mathematical counterpart of an experiment, is specified"—an entirely proper point of view for a mathematican to take. The second chapter, which covers conditional and compound probabilities, Bayes' formula and independence, will be found rather stiff reading to the uninitiated.

After a short discursus into elementary enumeration through the binomial and multinomial coefficients, with the arithmetic triangle and several combinatorial theories expounded, the remainder of the book is concerned with the properties of random variables. These are few brief notes towards the end on testing a statistical hypothesis and on decision-making.

For those persons who already know some probability theory this book will be enjoyable reading, if only to appreciate how familiar theorems can be wrapped up in mathematical language. For mathematicians in embryo who want to learn how a mathematician looks at the elements of probability theory this book will serve a useful educational purpose. For the statistician in embryo this book is not suitable since enlightenment can be achieved more simply elsewhere. (The reviewer has in mind, for example, *An Introduction to Probability Theory and its Applications* by W. Feller). But possibly the author did not intend to cater for this last class anyway.

ABSTRACTS

The following are abstracts of papers presented at meetings of the British Region held on October 27 and December 13, 1960.

745 W. T. WILLIAMS (Botany Department, University of Southampton). **Some Applications of Electronic Computers in Plant Ecology and Taxonomy.**

In ecological survey the population under study consists of sample areas (quadrats) defined by the presence or absence of a number of plant species, whose names and ecological requirements may not be known. The problem is to find the major discontinuities in the population. This can, in theory, be solved by successively subdividing the population so as to reduce the associations between the species by the greatest possible extent at each stage; this in turn involves setting up the correlation matrix at each stage and finding the species with the highest loading on the first principal axis of this matrix. To hold the entire matrix in a computer reduces unacceptably the space available for data. Instead of r, therefore, $|r|$ is calculated, and the $|r|$ values for each species summed as they are obtained. The species with the highest sum of $|r|$ is necessarily that with the highest loading on the first centroid axis of the matrix ($|r|$); this has been found to be an acceptable approximation to the answer required. Examples will be given of the application of this method to actual areas.

Valuable information on the structure of plant communities can be obtained by carrying out the corresponding inverse (Q-technique) subdivision; and, by collating the direct and inverse analyses, a completely objective account of the vegetation of a complex area can be given in extremely economical terms.

In taxonomy it is sometimes desirable to carry out a conventional factor anlysis, with communalities and rotation. No programmes of this type seem to be available in Britain; but the standard programmes for the extraction of latent roots and vectors greatly reduce computational labour. Moreover, the existence of general "matrix schemes" makes calculation of specification equations a relatively simple task.

746 J. N. R. JEFFERS (Forestry Commission, Wrecclesham, Farnham, Surrey). **Data Processing in Biological Research.**

The first task in the application of electronic digital computers to the statistical analysis of data arising in biological research has been the writing of the programmes for the actual analysis. This is, however, only a small part of the whole process, and the paper presents the results of recent work on the recording, compilation, and storage of data on paper tape and magnetic tape. This work suggests the need for a complete revision of the methods used in handling biological data, and has shown very large savings in the cost of research and in the unproductive use of key research workers. Also considered is the problem of the output results from the computer in a directly intelligible form, though progress has been slower in this field.

747 F. YATES (Rothamsted Experimental Station, Harpenden, Herts.). **Problems Arising in the Use of Electronic Computers in Statistical Analysis.**

A small electronic computer has been in use in the Rothamsted Statistics Department for the last six years. It has been used for a wide variety of statistical problems arising in agricultural and biological research. It was early applied to the analysis of replicated experiments and a large number of such analyses are now performed every year. More recently its use in the analysis of surveys has been developed. A general programme which has been written for survey analysis will be outlined and used to illustrate the types of problem that are encountered in using electronic computers for statistical work.

748 C. W. EMMENS (University of Sydney, N. S. W., Australia and Royal Veterinary College, London). **The Planning and Analysis of Some Field Trials with Cattle.**

Three field trials of the fertility of deep frozen semen have recently been conducted in Sydney. They were as far as possible completely balanced, on a factorial scheme, and have given remarkably consistent results, with residual χ^2-values within expected limits. This is attributed to the care with which randomisation and technique were controlled under difficult conditions.

749 C. C. SPICER (Imperial Cancer Res. Fund, Lincoln's Inn Fields, London). **Problems in the Analysis of a Large Scale Clinical Trial.**

The purpose of this trial was the evaluation of chemotherapy in the treatment and prevention of chronic bronchitis. It presents a number of points of interest both in planning and analysis.

The following are abstracts of papers presented at a special meeting on "BIOMETRICAL ASPECTS OF PLANT GROWTH" held in London on January 2, 1961, by the British Region in conjunction with the Society for Experimental Biology.

750 J. A. NELDER (National Vegetable Research Station, Wellesbourne, Warwick). **Models and Experiments for Growth Analysis.**

The relationship of growth analysis to other forms of statistical analysis, particularly those of long-term experiments and multivariate analysis, is indicated. The necessity for the development of methods for analysing the inter-related growths of different parts of a plant is stressed, and a first effort model of differential equations analagous to linear regression is proposed. Types of deviation from the assumed model, both systematic and random, and the kind of experiment needed to detect them are discussed together with problems arising in the estimation of parameters in growth equations. Possible applications of the model are indicated.

751 S. C. PEARCE and C. S. MOORE (East Malling Research Station, Nr. Maidstone, Kent). **A Study of the Sources of Variation in Growth of Fruit Trees.**

A method of studying growth previously described (*Biometrics 16*, 1–6) is applied to annual records of apple tree trunk circumference. The purpose of the investigation is to determine the main causes of variation in field trials with clonal fruit trees. Three standard errors are plotted against the logarithmic development of girth, namely, (a) log (girth), (b) increment in log (girth) since planting, and (c) log (girth) adjusted by girth at planting. A study of these curves enables the total variation resulting after a period of growth to be divided between initial variation, and initial growth rate negatively correlated with initial size differences, and a subsequent growth rate which is uncorrelated with initial size. Comparison of groups of trees having different treatments demonstrates the importance of good growing conditions in minimising variation. The reduction in positional variation by removing block effects and the results of cropping are also considered. The effects of these modifications in conditions accord with the theoretical model. Some measures are suggested for minimising error in field trials.

752 M. J. R. Healy (Rothamsted Experimental Station, Harpenden, Herts.). **Experiments for Comparing Growth Curves.**

Consider an experiment designed to study the differential effects of treatments on the growth curve of some organism. Suppose that the experimental subjects are allocated to the treatments by some process of randomisation and that observations are taken when the treatments are first applied and at regular intervals thereafter. Wishart suggested that tests of significance for the existence of treatment effects could be made by fitting a polynomial to the data from each subject and by applying standard statistical techniques to the coefficients of the polynomials. The paper discusses Wishart's technique and extends it to provide estimates of the treatment effects.

A general discussion was led by F. L. Milthorpe (University of Nottingham.)

CORRECTION

E. N. Hey and M. H. Hey [1960]. The Statistical Estimation of a Rectangular Hyperbola. *Biometrics 16*, 606–17.

The expression, nine lines from the bottom of page 609, should coincide with the column vector on the right of the matrix equation at the top of page 610. The solution for a, b, and $c - ab$ on page 610 should have coefficients bracketed to agree with the matrix equation at the top of the page. In footnote 2, page 615, Ey in the matrix should be E_y and on page 617, line 6, χ should be replaced by x.

THE BIOMETRIC SOCIETY

International

ANNUAL FINANCIAL STATEMENTS

The Biometric Society 1960

BALANCE SHEET

Assets		
Cash: Bank Balance	$9,605.90	
Petty Cash	19.40	$9,625.30
Liabilities		
Subscriptions, 1961	$ 107.50	
Dues, 1961	54.50	$ 162.00
Surplus, Jan. 1, 1960 (including petty cash)	7,048.64	
Gain for Period	2,414.66	9,463.30
		$9,625.30

Audited: B. G. Selvidge (Signed)
Date: March 20, 1961

INCOME AND EXPENDITURE STATEMENT

Income		
Subscriptions, 1959	$ 358.75	
Subscriptions, 1960	6,225.00	$ 6,583.75
Dues, 1959	105,50	
Dues, 1960	2,466.00	2,571.50
Sustaining memberships, 1960		650.00
Back dues and subscriptions		32.00
Regional allotments	116.50	
BIOMETRICS allotment from sustaining members	250.00	
Back issues	93.50	
Member subscriptions to Journal of ASA	35.00	
Sale of directories	2.00	
Overpayments	20.50	
Printing—Harvard University Press	30.00	
Advance to Pharmacology Symposium	1,000.00	
Bank refund of service charge	1.00	
	1,548.50	

Less: Credits and regional allotments used	138.00	1,410.50
Total Income		$11,247.75
Expenditures		
BIOMETRICS	$6,996.28	
Postage	250.23	
Office supplies	2.89	
Printing	260.29	
President's Office	200.00	
Member subscriptions to Journal of ASA	35.00	
Addressing and mailing services	88.40	
Pharmacology Symposium	1,000.00	$ 8,833.09
Excess of Income over Expenditures		2,414.66
		$11,247.75

Audited: B. G. Selvidge (Signed)
Date: March 20, 1961

SECRETARY'S ACCOUNT

STATEMENT OF INCOME AND EXPENDITURE

For the Period 1st January, 1959 to 31st December, 1960

Income	
Balance in hand, 31st December, 1958	£ 64.10. 4d
Received from Treasurer	266.11. 3d
Menbership dues received	1.12. –d
	£332.13. 7d
Expenditures	
Office equipment and stationary	£ 56. 2. 6d
Secretarial assistance	150. –. –d
Printing and binding	13.10. 7d
Postage	49.12. 1d
Travel	29.15. 9d
Books	18. 8. –d
Membership dues paid over	1.13. 6d
	£319. 2. 5d
Balance, 31st December, 1960	£ 13.11. 2d

I certify the above to be a true record of my transactions on behalf of the Biometric Society

Signed—....................
M. J. R. Healy
Secretary

I have examined the account book and vouchers produced by the Secretary and certify that the above statement is in accordance therewith.

Signed—....................
E. Church, A.A.C.C.A.

BIOMETRICS, VOLUME 16

STATEMENT OF OPERATIONS

For the Year Ending January 31, 1961

Income			
Biometric Society:			
907 Subscriptions at	$ 4.00	$ 3,628.00	
983 Subscriptions at	2.75	2,703.25	
9 Sustaining Members at	25.00	225.00	$ 6,556.25
1161 Direct Subscriptions at	7.00		8,127.00
Sale of Back Issues:			
Biometric Society		$ 374.00	
Editor's Office		4,474.00	4,848.00
March 1960 Issues at Cost to Biometric Society			61.28
Sale of Reprints			1,537.75
Over Payments			11.50
Payments to Regions			6.50
Total Income from Operations			$21,148.28
Expenditures			
Cost of Printing Journals			
Issue No. 1	$2,928.85		
Issue No. 2	3,344.71		
Issue No. 3	2,988.44		
Issue No. 4	4,467.72	$13,729.72	
Mailing and Express Charges			
Issue No. 1	$ 206.28		
Issue No. 2	229.82		
Issue No. 3	310.70		
Issue No. 4	346.56	1,093.36	$14,823.08
Cost of Printing Reprints (See Comments)			
Issue No. 4 (1959)	$ 290.69		
Issue No. 1	329.43		
Issue No. 2	331.79		
Issue No. 3	370.75	$ 1,322.66	
Mailing Charges Reprints (See Comments)			
Issue No. 4 (1959)	$ 24.01		
Issue No. 1	27.99		
Issue No. 2	22.75		
Issue No. 3	38.94	113.69	$ 1,436.35

Operating Expenses			
Stamps	$ 297.28		
Office Supplies	523.68		
Mailing Lists and Addressographing	131.75		
Back Issue Moving Costs	182.94		
Back Issue Handling and Storage	329.76		
Back Issue Insurance	46.35		
Auditing	45.00		
Bank Exchange and Service Charges	2.43		
Refunds and Overpayments on Subscriptions	55.25		
Salaries and F.I.C.A. Taxes	525.02		
Transfers to Regions	6.50		
Express Charges	13.97	$ 2,159.93	
Total Expenditures from Operations			18,419.36
Surplus from Operations			$ 2,728.92
Non-Operating Items			
Income			
Bank Interest	$ 680.75		
Refund on Insurance Policy	31.66		
Bank Credit and Exchange	2.06		
Announcement	57.00		
Refund on Blacksburg Postage Deposit	66.61		
Correction of Surplus from Previous Volumes (See Comments)	28.35	$ 866.43	
Expenditures			
Credits of 1959 Accounts Receivable		33.50	
Surplus from Non-Operating Items			832.93
Gross Surplus, 1-31-61			$ 3,561.85

BIOMETRICS, VOLUME 16
BALANCE SHEET
January 31, 1961

Assets		
Accounts Receivable	$ 2,089.25	
Bank Balances		
Savings Account #1 (Blacksburg)	10,397.02	
Savings Account #2 (Blacksburg)	9,637.50	
Savings Account #3 (Blacksburg)	3,168.41	
Checking Account (Tallahassee)	10,875.78	
Prepaid Insurance (Expires 4-25-63)	139.05	
Total Assets		$36,307.01

Liabilities and Surplus

Subscriptions to Volume #17	$ 5,993.75	
Subscriptions to Volume #18	182.00	
Subscriptions to Volume #19	42.00	
Subscriptions to Volume #20	7.00	
Surplus from Previous Volumes	26,520.41	
Surplus from Volume #16	3,561.85	
Total Liabilities and Surplus		$36,307.01

Comments

(1) Not included in Assets is stock of back issues from Volumes 1–16.

(2) Not included in Expenses is printing bill for December reprints of $507.20

(3) Corrections of surplus from previous volumes results from a reconciliation of subscription card files and actual liabilities due to prepaid subscriptions.

Audited February 17, 1961
Richard Q. Conrad (Signed)

Region Belge

L'Assemblée Générale annuelle de la Société Adolphe Quetelet s'est tenue à Bruxelles, dans la salle du Conseil de la Bibliothèque Royale de Belgique, le 25 février 1961. Le Conseil d'Administration pour 1961 et 1962 a été constitué comme suit:

Président: Professeur Maurice WELSCH.
Vice-Présidents: Prof. R. CONSAEL, H. LAUDELOUT, A. LECRENIER, J. REUSE et A. VAN DEN HENDE.
Secrétaire: Dr. L. MARTIN.
Secrétaire-adjointe: Anne LENGER.
Trésorier: Pierre GILBERT.
Membres: Dr. H. G. LION,
Mme OSLET-CONTER,
Mr. H. ROGGEN.

L'Assemblée Générale a été suivie d'une réunion organisée conjointement par la Société Quetelet, le Comité belge d'histoire des Sciences et la Société belge de Statistiquë Les exposés suivants ont été donnés:

"*Les rapports historiques entre l'anthropologie et la statistique*" par Monsieur A. LEGUEBE;

"*La distribution des caractères anthropométriques*" par le Professeur C. GINI, Président de l'Institut International de Sociologie à Rome.

British Region

At a meeting held on February 28th, 1961, the following papers were read and discussed:

P. D. Oldham—The Distribution of Arterial Pressure in the General Population.
L. R. Taylor—A Power Law Transformation for Aggregated Populations.

W.N.A.R.

TREASURER'S REPORT—DECEMBER 31, 1960

Bernice Brown, *Treasurer WNAR*

Receipts

Balance forwarded from previous treasurer, Dec., 1959	$ 461.22
Received from 1960 dues (143 @ 7, 10 @ 4)	
1961 dues (61 @ 7, 1 @ 4)	1472.00
Received for orders of back issues of *BIOMETRICS*	16.00
Received from Biometric Society (credit for one institutional membership)	10.00
Received from Biometric Society Treas. (credit for 4 transfers from ENAR)	4.00
Received from ENAR for information circulars	41.60
Total	$2004.82

Expenses

To Treasurer International (143 @ 6, 10 @ 4)	898.00
To *BIOMETRICS* for back issues ordered by members	16.00
To Marion Sandomire—mailing expense	11.15
To printers for stationary and supplies	111.84
To C. Zippin—partial expense to meeting	26.18
To W. Becker—mailing expense	12.00
To ENAR Treasurer—mailing expense	5.00
To printers for notices of meetings	45.40
To Pullman Herald for stationary	92.08
Total	$1217.65
On hand—checkbook balance	$ 787.17
Total	$2004.82

NEWS AND ANNOUNCEMENTS

Members are invited to transmit to their National or Regional Secretary (if members at large, to the General Secretary) news of appointments, distinctions, or retirements, and announcements of professional interest.

Robert W. Allard has been granted a Guggenheim Fellowship to study stochastic processes in genetics during a six months sabbatical leave to be spent at Oxford University in the Biometric Unit beginning February 1, 1961.

Neeti R. Bohidar has accepted the position of Assistant Professor of Statistics in the Department of Applied Statistics, Statistical Laboratory, Utah State University, Logan, Utah. He received his Ph.D. degree in Statistics from Iowa State University in 1960.

Joseph G. Bryan, formerly Senior Operations Research Analyst at the Central Research Laboratory, American Machine and Foundry Co., recently changed positions and is now employed by Travelers Weather Research Center, Inc., 11 Newport Avenue, West Hartford 7, Connecticut.

William R. Buckland has resigned from London Transport Executive to take charge of the new Statistical Advisory Service of The Economist Intelligence Unit Ltd., 22, Ryder Street, London, S.W.1.

Lyle D. Calvin, Oregon State University, spent the 1960 fall term on sabbatical leave with the Department of Statistics, Harvard University.

Douglas G. Chapman participated in the Fourth Annual Meeting of the North Pacific Fur Seal Commission held in Tokyo, January 20 to February 4 as a Scientific Advisor to the U. S. Commissioner. The North Pacific Fur Seal Commission is composed of representatives from Canada, Japan, U.S.S.R., and the U.S.A.

Franklin A. Graybill recently left Oklahoma State University to become Professor of Mathematical Statistics and Director of the Statistical Laboratory at Colorado State University, Fort Collins, Colorado.

Iowa State University granted Doctor of Philosophy degrees, with major in Statistics, to the following people during the Winter Quarter graduation: Roger McCullough, Jose Nieto de Pascual, J. N. K. Rao and Thomas Neil Throckmorton.

Gwilyn Jenkins, after spending a year at Stanford, is now back at his previous post as lecturer in Mathematical Statistics, University of London.

Kuo Hwa Lu, formerly Associate Professor, Department of Statistics, Utah State University, has taken the position of Associate Professor of Biostatistics at the University of Oregon Dental School, Portland.

Lincoln E. Moses, Professor of Statistics at Stanford University, is spending a year on sabbatical leave at the Department of Social Medicine, Oxford University, England. Dr. Moses received a Guggenheim Fellowship.

Donald B. Siniff had taken a position as Biometrician in the Division of Biological Research, Alaska Department of Fish and Game. He previously was at North Carolina State College working under the Southeastern Cooperative Wildlife Statistics Project.

TRAVEL GRANTS FOR ATTENDANCE AT THE INTERNATIONAL CONGRESS OF MATHEMATICIANS

Travel grants will be made to a number of mathematicians who wish to attend the International Congress of Mathematicians in Stockholm, on August 15-22, 1962. It is hoped that funds available through various sources may provide travel assistance for a considerable number of mathematicians.

There will be a greater effort than in the past to give aid to younger people. As grants will be made only to those who have filed applications, it is urgent that any who wish to receive a grant should fill out and file an application. Younger people are urged to file applications so that their cases can be considered. Applications can be obtained from the Division of Mathematics, National Academy of Sciences, National Research Council, Washington 25, D. C. by requesting an application for a travel grant to the 1962 International Congress.

The deadline for filing of applications is November 1, 1961, and an attempt will be made to announce the grants by January 1, 1962.

Awarding of grants will be made only to those persons whose applications have been received, in good order, by November 1. The selection will be made by a committee consisting of the regular Committee on Travel Grants of the Division of Mathematics of the National Academy of Sciences—National Research Council enlarged to include representatives of societies affiliated with the Division and representatives of various governmental agencies.

TRAINEESHIPS FOR PUBLIC HEALTH STATISTICIANS

The Public Health Service has announced the availability of traineeships for graduate training of professional public health personnel during the 1961-1962 academic year.

Traineeships in public health statistics are available to qualified persons. They provide stipends from a minimum of $250 per month for a post-bachelor candidate to a maximum of $400 per month for a post doctoral candidate and additional allowances for dependents, travel of the trainee, and academic tuition and fees.

Additional information and application forms may be secured from the Division of General Health Services, Public Health Service, U. S. Department of Health, Education, and Welfare, Washington 25, D. C.

GRADUATE TRAINEESHIPS IN BIOMETRY

Training programs designed to prepare students in the application of statistical and mathematical methods to biological problems, particularly those related to health and medical sciences, now exist in more than 20 universities throughout the country. Supported by training grant funds from the Public Health Service, NIH, these programs provide unusual opportunities for careers in teaching, research, and consultation. Employment opportunities for biometricians are excellent, with the demand by governmental and voluntary health agencies, medical research and educational institutions, and industry running far in excess of the available supply of trained personnel.

Programs of study are individually designed to lead to doctoral degrees, and in special instances, to other academic degrees. Traineeship stipends are provided at various levels depending on previous education and experience of the trainee and include allowances for dependents. Substantially full economic support or partial support may be provided, depending upon the proportion of time spent in training.

Interested applicants are encouraged to correspond with one of more of the Program Directors listed below because course offerings, as well as specific research problems for application of learned skills, vary from school to school.

Dr. Virgil Anderson
Purdue University
Lafayette, Indiana

Dr. George F. Badger
Western Reserve University
Cleveland 6, Ohio

Dr. Jacob E. Bearman
University of Minnesota
Minneapolis 14, Minnesota

Dr. Antonio Ciocco
University of Pittsburgh
Pittsburgh 13, Pennsylvania

Dr. Wilfrid J. Dixon
Medical Center, U.C.L.A.
Los Angeles 24, California

Dr. W. T. Federer
Cornell University
Ithaca, New York

Dr. John W. Fertig
Columbia University
New York 32, New York

Dr. Franklin A. Graybill
Colorado State University
Fort Collins, Colorado

Dr. Bernard G. Greenberg
University of North Carolina
Chapel Hill, North Carolina

Drs. John Gurland and T. A. Bancroft
Iowa State University
Ames, Iowa

Dr. Boyd Harshbarger
Virginia Polytechnic Institute
Blacksburg, Virginia

Dr. Allyn Kimball
Johns Hopkins University
Baltimore 5, Maryland

Dr. Schuyler G. Kohl
State Univ. of N. Y., Col. of Med.
Brooklyn 3, New York

Dr. Robert F. Lewis
Tulane University
New Orleans 12, Louisiana

Dr. Eugene Lukacs
The Catholic University of America
Washington, D. C.

Dr. Paul Meier
University of Chicago
Chicage 37, Illinois

Prof. Felix Moore
University of Michigan
Ann Arbor, Michigan

Dr. Lincoln E. Moses
Stanford Medical School
Stanford, California

Dr. Hugo Muench
Harvard School of Public Health
Boston 15, Massachusetts

Dr. Robert Quinn
Vanderbilt University
Nashville 5, Tennessee

Prof. J. A. Rigney
North Carolina State College
Raleigh, North Carolina

Drs. W. W. Schottstaedt and James Hagans
University of Oklahoma
Oklahoma City 4, Oklahoma

Dr. Malcolm E. Turner, Jr.
Medical College of Virginia
Richmond, Virginia

Dr. Colin White
Yale School of Medicine
New Haven, Connecticut

Dr. J. Yerushalmy
University of California
Berkeley 4, California

For those unable to train during the academic year, an unusual opportunity is provided by a cooperative GRADUATE SESSION OF STATISTICS IN THE HEALTH SCIENCES sponsored by these Program Directors and made possible by a training grant from the PHS, NIH. For information concerning available stipends and course offerings at elementary, intermediate, or advanced levels for the summers of 1961 and 1962, write Dr. Jacob E. Bearman, University of Minnesota, Minneapolis, Minnesota.

FELLOWSHIPS FOR MATHEMATICS GRADUATES

The Division of Preventive Medicine and the Institute of Industrial Health of the College of Medicine, The University of Cincinnati, with the support of the Public Health Service, are instituting a training program in Epidemiology and Biostatistics. As part of this program training and experience will be given in biology and research to bright, young, and prospective candidates for a Ph.D. degree in Mathematics or Statistics. The prospective fellow should have his B.A. or B.S. in Mathematics. He should have decided on a department at which he will want to complete his graduate work. Above all he must have demonstrated a level of achievement that will make him an acceptable candidate for an advanced degree at any university.

The student will be exposed also to research projects; working with medical and other fellows in this program and in the Institute of Industrial Health. He will be expected to write a publishable paper in some aspects of Mathematical Biology at the end of his two year stay. Arrangements may also be made for him to take additional work or gain additional research experience at another institution. The candidate will receive invaluable experience for a future career in the area of Biology or Mathematics and Statistics. The stipend will depend on qualifications and marital status and will range from $3400 to $4400 plus tuition.

For further information write to Dr. Theodor D. Sterling, Department of Preventive Medicine, College of Medicine, University of Cincinnati, Cincinnati, Ohio.

ON-THE-JOB TRAINEE APPOINTMENTS

The Department of Statistics at Roswell Park Memorial Institute has been awarded a grant by the National Cancer Institute and in conjunction with the University of Buffalo offers a two-year graduate on-the-job training program leading to a Master of Science degree in biostatistics.

Students with a Bachelor's degree, and possessing a good background in mathematics (through calculus), and course work in physical, biological or behavioral sciences who will begin their graduate work next fall are sought as applicants for these appointments. In addition to taking graduate course work, trainees will receive on-the-job training in research studies that are currently in progress at Roswell Park Memorial Institute. Trainees will receive an average annual stipend of $2300.00 plus tuition.

Requests for further information and application forms should be addressed to Miss Barbara A. Cote, c/o Dr. Irwin D. J. Bross, Chief, Department of Statistics, Roswell Park Memorial Institute, 666 Elm Street. Buffalo 3, New York.

BIOMETRIE—PRAXIMETRIE

Tome I, N° 3–4 *Juillet-Octobre, 1960*

Tome II, N° 1 *Mars, 1961*

Publishers for the

SOCIETE ADOLPHE QUETELET
7 rue Héger-Bordet, Brussels 1, Belgium

Presses Universitaires de Bruxelles
50, av. Franklin Roosevelt
Brussels

The Biometric Society

BIOMETRICS

FOUNDED BY THE BIOMETRICS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION

TABLE OF CONTENTS

Articles

Queries and Notes

Book Reviews

Number 3 September 1961 Volume 17

Officers of the Biometric Society

General Officers

President: Leopold Martin

Secretary: M. J. R. Healy; *Treasurer*: M. A. Kastenbaum

Council Members

1959–1961	1960–1962	1961–1963
M. S. Bartlett, *BR*	G. S. Watson, *ENAR*	A. W. Kimball, *ENAR*
D. G. Chapman, *WNAR*	C. I. Bliss, *ENAR*	H. N. Turner, *AR*
C. W. Emmens, *AR*	D. J. Finney, *BR*	S. C. Pearce, *BR*
F. G. Fraga, *R. Bras.*	A. Linder, *Switzerland*	J. L. Hodges, *WNAR*
A. Lenger, *R. Belg.*	P. V. Sukhatme, *R. Ital.*	W. U. Behrens, *DR*
C. C. Li, *ENAR*	G. Teissier, *RF*	H. L. LeRoy, *Switzerland*
	F. Yates, *BR*	

Regional Officers

Region	*President*	*Secretary*	*Treasurer*
Australasian	M. Belz	W. B. Hall	G. W. Rogerson
Belgian	M. Welsch	L. Martin	P. Gilbert
Brazilian	A. M. Penha	E. Berquó	A. Groszmann
British	J. A. Fraser Roberts	C. D. Kemp	P. A. Young
E. N. American	Oscar Kempthorne (Henry L. Lucas)	Erwin L. LeClerg	Donald A. Gardiner
French	Ph. L'Heritier	Sully Ledermann	Sully Ledermann
German	O. Heinisch	R. Wette	M. P. Geppert
Italian	G. Montalenti	R. Scossiroli	F. Sella
W. N. American	W. A. Taylor	W. A. Becker	Bernice Brown

National Secretaries

Denmark	N. F. Gjeddebæk	*Netherlands*	H. de Jonge
India	A. R. Roy	*Norway*	L. K. Strand
Japan	M. Hatamura	*Sweden*	H. A. O. Wold
	Switzerland	H. L. LeRoy	

THE GROWTH AND AGE-DISTRIBUTION OF A POPULATION OF INSECTS UNDER UNIFORM CONDITIONS

E. J. Williams
Division of Mathematical Statistics,
C.S.I.R.O., Canberra, Australia.

I. INTRODUCTION

This paper deals with the growth of populations of insects such as aphids that reproduce continuously under uniform conditions. Its object is to derive the relations among the different parameters of the populations: the birth rate, the death rates at different stages, the age-distribution, and the intrinsic rate of natural increase (Lotka [1945]). Interest in these relations arose from the need to estimate the size of aphid populations by indirect means; the work presented here is supplementary to that of Hughes [1961], who was responsible for initiating the experimental investigations and who saw the possibility of estimating the rate of growth of an aphid population by means of a study of its age-distribution.

Insects pass through a number of immature stages (instars) which may be readily distinguished. If the average duration of each instar is known, a count of the numbers of insects in each provides a conveniently grouped age-distribution. By studying this age-distribution at successive intervals we can estimate the way in which the population is developing and thus indirectly determine the rate of growth of the population.

Although the original investigation was of an aphid population, the problem will be framed in terms sufficiently general to apply to several types of insect with similar pattern of development.

II. FORMULATION OF THE PROBLEM

We shall set out in rather general terms the conditions under which the population is developing, so that the results may be applied to different types of insect.

We consider an insect with k instars, and designate the parameters of the population as follows:

γ = birth rate of adults[1],
μ_i = death rate during the ith instar,
μ_A = death rate of adults.

For simplicity these parameters will all be assumed constant within the stage of growth (immature or adult) to which they apply.

We denote by $n_i(t)$ the number in the ith instar at time t, and by $n_A(t)$ the number of adults at time t.

We shall assume that the durations of successive instars are independently distributed; although it could have been assumed that the lengths of successive instars for any insect are negatively correlated, so that an instar of unusually long duration will be partly compensated by a short one, the evidence for this assumption is not convincing.

We shall denote by x_i the duration of the ith instar for some insect, and by X_i the duration of the first i instars, so that

$$X_i = x_1 + \cdots + x_i .$$

Both x_i and X_i are random variables, whose probability densities we shall denote by $f_i(x_i)$ and $f^i(X_i)$, respectively. Under the assumption of independence, the density $f^i(X_i)$ is the i-fold convolution of the densities for the first i instars.

We denote the corresponding distribution functions by F with appropriate affixes; thus

$$F_i(x_i) = \int_0^{x_i} f_i(u)\, du,$$

and

$$F^i(X_i) = \int_0^{X_i} f^i(u)\, du.$$

Since under uniform conditions the population will be increasing or decreasing at a constant rate, we shall find it convenient to express the population size in terms of the Laplace transforms of the various probability densities. We shall write

$$\phi_i(\tau) = \int_0^\infty e^{-\tau u} f_i(u)\, du.$$

The Laplace transforms for the convolution distributions are simply products of those for successive instar distributions.

The Laplace transform plays an important part in many biological

[1] This is to be distinguished from the birth rate of the population as a whole, which is customarily denoted by λ (see Hughes [1961]).

problems. An interpretation of its biological applications has been given by Cox [1957].

The distribution of durations among survivors in any instar will be affected by the death rate. For the ith instar with death rate μ_i, the density of a duration x_i will be proportional to

$$e^{-\mu_i x_i} f_i(x_i)$$

and will thus be given by

$$g_i(x_i) = e^{-\mu_i x_i} f_i(x_i)/\phi_i(\mu_i).$$

We shall now consider the age-distribution, and derive thence the rate of growth of the population. In these populations under favourable conditions, the rate of increase is high and mortality has only a small effect. We therefore first consider how the population would grow if there were no mortality.

III. POPULATION GROWTH WITH NO MORTALITY

The chance that an insect is in the ith instar at a time t after birth is

$$P(X_i \geq t > X_{i-1}) = F^{i-1}(t) - F^i(t);$$

likewise, the chance that it is an adult is

$$P(t > X_k) = F^k(t).$$

Now the number of births in an interval $(u, u + du)$ is

$$\gamma n_A(u)\, du + o(du),$$

so the number expected in the ith instar at time t is

$$n_i(t) = \gamma \int_{-\infty}^{t} n_A(u)[F^{i-1}(t - u) - F^i(t - u)]\, du.$$

We have in particular for the adult stage the result

$$n_A(t) = \gamma \int_{-\infty}^{t} n_A(u)\, F^k(t - u)\, du. \tag{1}$$

Being of homogeneous type, the integral equation (1) has the solution

$$n_A(t) = n_A(0)e^{\rho t},$$

where ρ is the solution of the equation

$$e^{\rho t} = \gamma \int_{-\infty}^{t} e^{\rho u}\, F^k(t - u)\, du,$$

or

$$\frac{1}{\gamma} = \int_0^{\infty} e^{-\rho u}\, F^k(u)\, du = \frac{1}{\rho}\, \phi_1\phi_2\phi_3 \cdots \phi_k\,, \tag{2}$$

where we have suppressed the argument of the Laplace transforms. This equation corresponds to the one given by Lotka [1925], p. 115, equation (25).

In equation (2) the function $F^k(u)$ represents the probability that an individual is a member of the adult population at a time u after birth. It thus corresponds to Lotka's $p(a)$ (the probability of survival at time a after birth), although $p(a)$ is a decreasing function of a, whereas $F^k(X)$ is an increasing function of X. The modes of development of the two populations are quite different: the individuals of one are subject to mortality, whereas the individuals of the other are subject to delay in reaching maturity.

We can express the numbers of individuals in each instar in terms of the Laplace transforms with argument ρ.

$$n_i(t) = \gamma n_A(t) \int_0^\infty e^{-\rho u}[F^{i-1}(u) - F^i(u)]\,du$$

$$= \frac{\gamma}{\rho} n_A(0)e^{\rho t}\phi_1\phi_2 \cdots \phi_{i-1}(1 - \phi_i).$$

The total number of all stages at time t is

$$\frac{\gamma}{\rho} n_A(0)e^{\rho t}. \tag{3}$$

The fraction of the population in the ith instar is

$$\phi_1\phi_2 \cdots \phi_{i-1}(1 - \phi_i), \tag{4}$$

and the fraction in the adult stage is

$$\phi_1\phi_2 \cdots \phi_k = \frac{\rho}{\gamma} \tag{4A}$$

independently of t.

This exponential solution agrees with physical considerations, and describes the steadily increasing population to be expected when conditions have been constant for a long period. The development of the population is described by the constant growth rate ρ.

If, as a result of changing conditions, the initial age-distribution is different from that given by the solutions (4), the solution for the distribution between instars will be more complicated. However, it is convenient in practice to confine observations to steadily increasing populations under uniform conditions; the behaviour of such populations is fairly easily interpreted.

We can also find the distribution of durations in any instar (the 'instar age-distribution'), which is of some theoretical interest. In the steadily increasing population the number of individuals entering the ith instar a time x previous to observation is proportional to

$$e^{-\rho x}.$$

The number of individuals of instar-age x in the ith instar is the number which have entered a time x ago, reduced by the factor

$$1 - F_i(x),$$

for those which have left the ith instar. Hence the probability density of instar-ages in the ith instar is

$$\frac{e^{-\rho x}[1 - F_i(x)]}{\int_0^\infty e^{-\rho u}[1 - F_i(u)]\,du} = \frac{\rho e^{-\rho x}[1 - F_i(x)]}{1 - \phi_i(\rho)} \tag{5}$$

and the density of adult ages is

$$\rho e^{-\rho x}. \tag{5A}$$

The expressions (4) and (5) between them completely specify the age-distribution in a steadily increasing population without mortality. If we know the age-distribution between instars, we can in principle equate the expressions (4) to the known probabilities to determine the rate of growth of the population.

IV. EFFECT OF MORTALITY ON POPULATION GROWTH

When the effect of mortality is taken into account, we have to allow for the chance that the individual will die before reaching the instar considered. Also, since we are assuming a different mortality in each developmental stage, a somewhat more elaborate analysis is required than was given in Section III.

We assume that the net rate of growth of the population is ρ, and consider the movement of individuals through the ith instar. Since the death rate is μ_i, the number of individuals entering the ith instar at a time x previous to observation and surviving is proportional to

$$e^{-(\rho+\mu_i)x},$$

so that, as in Section III, the density of instar-ages is

$$\frac{(\rho + \mu_i)e^{-(\rho+\mu_i)x}[1 - F_i(x)]}{1 - \phi_i(\rho + \mu_i)} \tag{6}$$

and the density of adult ages is

$$(\rho + \mu_A)e^{-(\rho+\mu_A)x}. \tag{6A}$$

The effect of mortality in the ith instar is thus to replace ρ by $\rho + \mu_i$ in the distribution of instar-ages. In general it can be seen that the effect of mortality is to replace ρ by $\rho + \mu_i$ in all functions of the ith instar distribution.

The number of individuals going from instar $i - 1$ to instar i in an interval $(t, t + dt)$ is

$$\frac{n_{i-1}(t)\, dt(\rho + \mu_{i-1})}{1 - \phi_{i-1}(\rho + \mu_{i-1})} \int_0^\infty e^{-(\rho+\mu_{i-1})u} f_{i-1}(u)\, du + o(dt)$$

$$= n_{i-1}(t)\, dt \frac{(\rho + \mu_{i-1})\phi_{i-1}(\rho + \mu_{i-1})}{1 - \phi_{i-1}(\rho + \mu_{i-1})} + o(dt).$$

Similarly the number going from instar i to instar $i + 1$ is

$$n_i(t)\, dt \frac{(\rho + \mu_i)\phi_i}{1 - \phi_i} + o(dt),$$

where the argument of the Laplace transform has been suppressed since it is indicated by the subscript.
Also the number dying is

$$n_i(t)\, dt\mu_i + o(dt).$$

Hence, since the rate of population growth is ρ, we have

$$n_i(t) \frac{\rho + \mu_i}{1 - \phi_i} = n_{i-1}(t) \frac{(\rho + \mu_{i-1})\phi_{i-1}}{1 - \phi_{i-1}}, \tag{7}$$

and

$$n_A(t)(\rho + \mu_A) = n_k(t) \frac{(\rho + \mu_k)\phi_k}{1 - \phi_k}. \tag{7A}$$

Taking into account the birth rate γ, we have

$$n_1(t) \frac{(\rho + \mu_1)}{1 - \phi_1} = \gamma n_A(t).$$

Thus the numbers in each instar can be expressed in terms of the number of adults living at the same time:

$$n_i(t) = \gamma n_A(t) \frac{\phi_1\phi_2 \cdots \phi_{i-1}(1 - \phi_i)}{\rho + \mu_i}. \tag{8}$$

Similarly, from (7A) we have

$$n_A(t) = \gamma n_A(t) \frac{\phi_1\phi_2 \cdots \phi_k}{\rho + \mu_A},$$

giving the equation satisfied by the growth rate:

$$\frac{\rho + \mu_A}{\gamma} = \phi_1\phi_2 \cdots \phi_k , \tag{9}$$

which corresponds to (2).

It is convenient to express equation (9) explicitly as an integral equation similar to (2). In this integral equation the densities g_i of durations adjusted for death rate replace the original densities f_i .

The basic relation for this representation is

$$\begin{aligned}\phi_i(\rho + \mu_i) &= \int_0^\infty e^{-(\rho+\mu_i)u} f_i(u)\, du,\\ &= \phi_i(\mu_i) \int_0^\infty e^{-\rho u} g_i(u)\, du,\\ &= \phi_i(\mu_i)\psi_i(\rho),\end{aligned}$$

where $\psi_i(\rho)$ is the Laplace transform of the density $g_i(x_i)$. This relation expresses the Laplace transforms appearing in (9) as the product of two transforms, one depending on the death rate for the instar, the other on the growth rate of the population.

Thus equation (9) is explicitly

$$\begin{aligned}\frac{\rho + \mu_A}{\gamma} &= \phi_1(\rho + \mu_1)\phi_2(\rho + \mu_2) \cdots \phi_k(\rho + \mu_k)\\ &= \phi_1(\mu_1)\phi_2(\mu_2) \cdots \phi_k(\mu_k)\psi_1(\rho)\psi_2(\rho) \cdots \psi_k(\rho) \qquad (10)\\ &= \phi_1(\mu_1)\phi_2(\mu_2) \cdots \phi_k(\mu_k) \int_0^\infty e^{-\rho u} g^k(u)\, du.\end{aligned}$$

Finally, equation (10) can be expressed in a manner similar to Lotka's equation if we observe that, $\psi_A(\rho) = \mu_A/(\rho + \mu_A)$.

The equation then becomes

$$\frac{\mu_A}{\gamma} = \phi_1(\mu_1)\phi_2(\mu_2) \cdots \phi_k(\mu_k) \int_0^\infty e^{-\rho u} g^A(u)\, du, \tag{11}$$

where $g^A(X_A)$ is the distribution of the total life span. Equation (2) clearly cannot be expressed in this way, since (11) would be meaningless if there were no mortality.

The reason why equations (10) and (11) differ from the original Lotka equation is that we have here assumed that the death rate is not directly time-dependent, but instar-dependent. Thus the rate of growth of the population depends on the distribution of durations of the different instars. When the death rate in each instar is the same ($\mu_i = \mu$), equation (10) may be reduced to

$$\frac{1}{\gamma} = \int_0^\infty e^{-(\rho+\mu)u} F^k(u)\, du,$$

which is comparable with (2) and with Lotka's original equation.

V. RESULTS FOR PARTICULAR INSTAR DURATION DISTRIBUTIONS

The results of Sections III and IV are intractable unless some assumption is made about the distribution of the durations of the instars. We shall here consider two simple distributions which may be suitable approximations and may indicate the form that the results will take in general. We give the results when there is mortality, from which the results for no mortality can be easily deduced.

(a) *Constant instar length*

If the length of the *i*th instar is a constant χ_i , the results take a fairly simple form.

Firstly,

$$\phi_i(\tau) = e^{-\tau\chi_i}.$$

The probability density of instar-ages in the *i*th instar

$$= \frac{(\rho + \mu_i)e^{-(\rho+\mu_i)x}}{1 - e^{-(\rho+\mu_i)\chi_i}} \qquad (x \leq \chi_i),$$

$$= 0 \qquad (x > \chi_i).$$

The growth rate of the population is defined by

$$\frac{\rho + \mu_A}{\gamma} = \exp\left\{-\sum_{h=1}^{k} (\rho + \mu_h)\chi_h\right\},$$

and the number in the *i*th instar is

$$n_i(t) = \gamma n_A(t)\, \frac{(1 - e^{-(\rho+\mu_i)\chi_i}) \exp\left\{-\sum_{h=1}^{i-1} (\rho + \mu_h)\chi_h\right\}}{\rho + \mu_i}.$$

It is often reasonable to assume that death rate is independent of age. When this is done, the equations reduce to a simple form, which is found to give results corresponding closely with observation for many aphid populations.

Putting $\mu_i = \mu$, we have

$$n_i(t) = \gamma n_A(t)\, \frac{[1 - e^{-(\rho+\mu)\chi_i}] \exp\left\{-(\rho + \mu) \sum_{h=1}^{i-1} \chi_h\right\}}{\rho + \mu}.$$

The growth rate of the population is given by

$$\frac{\rho + \mu}{\gamma} = \exp\left\{-(\rho + \mu) \sum_{h=1}^{k} \chi_h\right\}.$$

It is seen that, if the death rate is increased, the actual growth rate is decreased by the same amount.

Hughes [1961] has worked with the assumption of constant death rate. He defines the sum $\rho + \mu$ as the potential rate of natural increase of the populations. The justification for this definition is that, if all causes of death could be removed, the rate of growth of the population would in fact be $\rho + \mu$.

(b) *Exponential distribution*

We here assume that the duration of the ith instar is exponentially distributed with mean χ_i;

$$f_i(x_i) = \frac{1}{\chi_i} e^{-x_i/\chi_i}.$$

The Laplace transform here takes the simple form

$$\phi_i(\tau) = 1/(1 + \tau\chi_i).$$

Then the distribution of instar-ages is given by

$$\frac{1 + (\rho + \mu_i)\chi_i}{\chi_i} e^{-(\rho+\mu_i+1/\chi_i)x},$$

which is of the same form as the distribution of durations, but with mean

$$\frac{\chi_i}{1 + (\rho + \mu_i)\chi_i}.$$

The growth rate of the population is given by

$$\frac{\rho + \mu_A}{\gamma} = \frac{1}{\prod_{h=1}^{k} [1 + (\rho + \mu_h)\chi_h]},$$

an algebraic equation of degree $k + 1$. The number in the ith instar is

$$n_i(t) = \frac{\gamma n_A(t)\chi_i}{\prod_{h=1}^{i-1} [1 + (\rho + \mu_h)\chi_h]}.$$

VI. REFERENCES

Cox, D. R. [1957]. Discussion on distributions associated with random walk and recurrent events. *J. Roy. Stat. Soc. B 19*, 113–4.

Hughes, R. D. [1961]. Factors in the biological control of aphids. I. Population parameters and their estimation (submitted for publication).

Lotka, A. J. [1925]. *Elements of Physical Biology.* Chapter IX. Baltimore: Williams and Wilkins.

Lotka, A. J. [1945]. Population analysis as a chapter in the mathematical theory of evolution. *In* Le Gros Clark, W. E., and P. B. Medawar, *Essays in Growth and Form,* 355–85. Oxford.

EMPIRICAL SAMPLING ESTIMATES OF GENETIC CORRELATIONS

L. D. VanVleck and C. R. Henderson
Cornell University, Ithaca, New York, U.S.A.

Two methods of estimating the genetic correlation between two traits are commonly used. One procedure utilizes estimates of within subclass covariances as described by Hazel [1943]. An example is given by the analysis of observations on parent-offspring pairs. The other method uses estimates of the group variance and covariance components for two characters from an analysis within and between groups of relatives. The between and within analysis of sib or half-sib groups is an example of this type of analysis. The sampling variances of the estimates obtained in either of these ways has not been widely investigated. Robertson [1959] has derived an estimate of the variance of the genetic correlation estimated from the variance and covariance analysis of groups of relatives with observations on two variables. His procedure deals with the special case in which both traits have the same heritability and in which all subclass numbers are equal. Tallis [1959] has given a general solution when subclass numbers are equal for estimates obtained from a between and within analysis of related groups. A general solution is, also, described by Mode and Robinson [1959] for genetic and genotypic correlations estimated from components of variance estimated from a four-way nested classification random model for the equal subclass numbers situation. For the parent-offspring method of estimation Reeve [1955] has given approximate formulae for the variance of the estimate. Apparently none of these procedures has been tested by empirical sampling. The purpose of this paper is to describe a procedure for obtaining empirical sampling estimates of genetic correlations in an attempt to learn something of the sampling variances of the estimates obtained from the parent-offspring analysis. Sampling variances of the empirical sampling estimates are then compared with the theoretical variances derived by Reeve [1955].

SAMPLING PROCEDURE

Let us first consider a sampling scheme for the one-way classification model which may be extended easily to more complex situations.

Suppose the model we are sampling is

$$Y_{ij} = \mu + \alpha_i + \delta_{ij} \qquad (i = 1, \cdots, c, j = 1, \cdots, n_i),$$

where

μ is the underlying population mean,
α_i is a random effect associated with the ith class, and
δ_{ij} is a random effect associated with the jth observation in the ith class.

Further assume that the α_i and δ_{ij} are NID $(0, \sigma_\alpha^2)$ and NID $(0, \sigma_\delta^2)$, respectively. Then the variance of Y_{ij} is $\sigma_\alpha^2 + \sigma_\delta^2$. To generate samples from such a population we need to draw samples from a NID (0, 1) population, multiply these random normal deviates by the appropriate standard deviations, and add to these products the population mean. Now $Y_{ji} = \mu + \sigma_\alpha a_i + \sigma_\delta e_{ij}$, where the a_i and e_{ij} are NID (0, 1). If we take $Y'_{ji} = Y_{ij} - \mu$, we have $Y'_{ij} = \sigma_\alpha a_i + \sigma_\delta e_{ij}$. The sums of squares normally computed in terms of our sampling model are:

$$\sum_{i=1}^{c} \sum_{j=1}^{n_i} Y'^2_{ij} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\sigma_\alpha a_i + \sigma_\delta e_{ij})^2$$

$$= \sigma_\alpha^2 \sum_{i=1}^{c} n_i a_i^2 + 2\sigma_\alpha \sigma_\delta \sum_{i=1}^{c} a_i e_{i.} + \sigma_\delta^2 \sum_{i=1}^{c} \sum_{j=1}^{n_i} e_{ij}^2 ,$$

$$\sum_{i=1}^{c} \frac{(Y'_{i.})^2}{n_i} = \sigma_\alpha^2 \sum_{i=1}^{c} n_i a_i^2 + 2\sigma_\alpha \sigma_\delta \sum_{i=1}^{c} a_i e_{i.} + \sigma_\delta^2 \sum_{i=1}^{c} \frac{e_{i.}^2}{n_i} ,$$

and

$$\frac{Y^{12}_{..}}{n_.} = \frac{1}{n_.} \left[\sigma_\alpha^2 \left(\sum_{i=1}^{c} n_i a_i \right)^2 + 2\sigma_\alpha \sigma_\delta \left(\sum_{i=1}^{c} n_i a_i \right) (e_{..}) + \sigma_\delta^2 e_{..}^2 \right].$$

(The usual dot notation signifies summation over that subscript.)

It is apparent that in order to generate a sample with given subclass numbers it is necessary to compute only six terms

$$\left(\sum_{i=1}^{c} n_i a_i^2 , \quad \sum_{i=1}^{c} a_i e_{i.} , \quad \sum_{i=1}^{c} \sum_{j=1}^{n_i} e_{ij}^2 , \quad \sum_{i=1}^{c} \frac{e_{i.}^2}{n_i} , \quad \sum_{i=1}^{c} n_i a_i , \quad \text{and} \quad e_{..} \right)$$

which are functions of the random normal deviates and subclass numbers. In order to complete the computation of the sampling sums of squares these functions can be multiplied by the corresponding constants which are σ_α^2, σ_δ^2, and $2\sigma_\alpha\sigma_\delta$ and added accordingly. Thus functions of deviates and subclass numbers can be generated and different sets of parameter values used with them in order to determine the effect of different ratios of the population variances on the sampling estimates of the variance components. This procedure can substantially reduce the amount of sampling time required. It should be noted that such estimates will be correlated, but usually this will not be a disadvantage and actually may be of interest. Extension of such a procedure to a multiple classification is apparent. For the two-way classification with interaction 28 terms would be generated.

A similar procedure may be used to generate estimates of genetic correlations. We are actually concerned with a four-variate model (2 traits on the parent and 2 on the offspring). For four-tuple samples of size N let the four-variate sampling model be:

$$\begin{aligned} X_{1i} &= \lambda_1 e_{1i} , \\ X_{2i} &= \lambda_2 e_{1i} + \lambda_3 e_{2i} , \\ X_{3i} &= \lambda_4 e_{1i} + \lambda_5 e_{2i} + \lambda_6 e_{3i} , \\ X_{4i} &= \lambda_7 e_{1i} + \lambda_8 e_{2i} + \lambda_9 e_{3i} + \lambda_{10} e_{4i} ; \end{aligned}$$

where the e's are random normal deviates NID (0, 1) and the λ's are constants determined by the variance-covariance matrix of the X's,

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}.$$

The X's can be arranged in such a way that two of them refer to a pair of traits on a parent and two of them to the same pair of traits on its offspring. In this paper let us consider the situation described by Table 1.

TABLE 1
ASSIGNMENT OF VARIABLES

Trait	Parent	Offspring
I	X_1	X_3
II	X_2	X_4

According to our model:

$$\sigma_1^2 = \lambda_1^2 , \qquad \sigma_3^2 = \lambda_4^2 + \lambda_5^2 + \lambda_6^2 ,$$
$$\sigma_2^2 = \lambda_2^2 + \lambda_3^2 , \qquad \sigma_4^2 = \lambda_7^2 + \lambda_8^2 + \lambda_9^2 + \lambda_{10}^2 ,$$

$$\begin{aligned}
\sigma_{12} &= \lambda_1\lambda_2 \ , & \sigma_{23} &= \lambda_2\lambda_4 + \lambda_3\lambda_5 \ , \\
\sigma_{13} &= \lambda_1\lambda_4 \ , & \sigma_{24} &= \lambda_2\lambda_7 + \lambda_3\lambda_8 \ , \\
\sigma_{14} &= \lambda_1\lambda_7 \ , & \sigma_{34} &= \lambda_4\lambda_7 + \lambda_5\lambda_8 + \lambda_6\lambda_9 \ .
\end{aligned}$$

Solving for the λ's in terms of the variances and covariances we obtain:

$$\begin{aligned}
\lambda_1 &= \sigma_1 \ , \\
\lambda_2 &= \sigma_{12}/\sigma_1 \ , \\
\lambda_3 &= [\sigma_2^2 - (\sigma_{12}^2/\sigma_1^2)]^{1/2}, \\
\lambda_4 &= \sigma_{13}/\sigma_1 \ , \\
\lambda_5 &= (\sigma_{23} - \sigma_{12}\sigma_{13}/\sigma_1^2)/[\sigma_2^2 - (\sigma_{12}^2/\sigma_1^2)]^{1/2} \\
\lambda_6 &= [\sigma_3^2 - (\sigma_{13}^2/\sigma_1^2) - (\sigma_{23} - \sigma_{12}\sigma_{13}/\sigma_1^2)^2/(\sigma_2^2 - (\sigma_{12})^2/\sigma_1^2)]^{1/2} \\
\lambda_7 &= \sigma_{14}/\sigma_1 \ , \\
\lambda_8 &= (\sigma_{24} - \sigma_{12}\sigma_{14}/\sigma_1^2)/[\sigma_2^2 - (\sigma_{12})^2/\sigma_1^2]^{1/2} \\
\lambda_9 &= \{\sigma_{34} - \sigma_{13}\sigma_{14}/\sigma_1^2 - (\sigma_{23} - \sigma_{12}\sigma_{13}/\sigma_1^2) \\
&\qquad \cdot(\sigma_{24} - \sigma_{12}\sigma_{14}/\sigma_1^2)/[\sigma_2^2 - (\sigma_{12})^2/\sigma_1^2]\}/\{\sigma_3^2 - (\sigma_{13})^2/\sigma_1^2 \\
&\qquad - (\sigma_{23} - \sigma_{12}\sigma_{13}/\sigma_1^2)^2/[\sigma_2^2 - (\sigma_{12})^2/\sigma_1^2]\}^{1/2},
\end{aligned}$$

and

$$\begin{aligned}
\lambda_{10} &= \langle\sigma_4^2 - (\sigma_{14})^2/\sigma_1^2 - (\sigma_{24} - \sigma_{12}\sigma_{14}/\sigma_1^2)^2/[\sigma_2^2 - (\sigma_{12})^2/\sigma_1^2] \\
&\qquad - \{\sigma_{34} - \sigma_{13}\sigma_{14}/\sigma_1^2 - (\sigma_{23} - \sigma_{12}\sigma_{13}/\sigma_1^2) \\
&\qquad \cdot(\sigma_{24} - \sigma_{12}\sigma_{14}/\sigma_1^2)/[\sigma_2^2 - (\sigma_{12})^2/\sigma_1^2]\}^2/[\sigma_3^2 - (\sigma_{13})^2/\sigma_1^2 \\
&\qquad - (\sigma_{23} - \sigma_{12}\sigma_{13}/\sigma_1^2)^2/[\sigma_2^2 - (\sigma_{12})^2/\sigma_1^2]\rangle^{1/2}.
\end{aligned}$$

Next let us define the following quadratics of the normal deviates:

$$\begin{aligned}
L_1 &= \sum_{i=1}^{N} e_{1i}^2 - \frac{e_{1.}^2}{N} \ , & L_{13} &= \sum_{i=1}^{N} e_{1i}e_{3i} - \frac{e_{1.}e_{3.}}{N} \ , \\
L_2 &= \sum_{i=1}^{N} e_{2i}^2 - \frac{e_{2.}^2}{N} \ , & L_{14} &= \sum_{i=1}^{N} e_{1i}e_{4i} - \frac{e_{1.}e_{4.}}{N} \ , \\
L_3 &= \sum_{i=1}^{N} e_{3i}^2 - \frac{e_{3.}^2}{N} \ , & L_{23} &= \sum_{i=1}^{N} e_{2i}e_{3i} - \frac{e_{2.}e_{3.}}{N} \ , \\
L_4 &= \sum_{i=1}^{N} e_{4i}^2 - \frac{e_{4.}^2}{N} \ , & L_{24} &= \sum_{i=1}^{N} e_{2i}e_{4i} - \frac{e_{2.}e_{4.}}{N} \ , \\
L_{12} &= \sum_{i=1}^{N} e_{1i}e_{2i} - \frac{e_{1.}e_{2.}}{N} \ , & L_{34} &= \sum_{i=1}^{N} e_{3i}e_{4i} - \frac{e_{3.}e_{4.}}{N} \ .
\end{aligned}$$

Using these definitions and the procedure described previously, the empirical sampling variance and covariance estimates with $(N - 1)$ degrees of freedom are the following functions of λ's, normal deviates, and N:

$$\hat{\sigma}_1^2 = (\lambda_1^2 L_1)/(N - 1),$$

$$\hat{\sigma}_2^2 = (\lambda_2^2 L_1 + \lambda_3^2 L_2 + 2\lambda_2\lambda_3 L_{12})/(N - 1),$$

$$\hat{\sigma}_3^2 = [\lambda_4^2 L_1 + \lambda_5^2 L_2 + \lambda_6^2 L_3 + 2(\lambda_4\lambda_5 L_{12} + \lambda_4\lambda_6 L_{13} + \lambda_5\lambda_6 L_{23})]/(N - 1),$$

$$\hat{\sigma}_4^2 = [\lambda_7^2 L_1 + \lambda_8^2 L_2 + \lambda_9^2 L_3 + \lambda_{10}^2 L_4 + 2(\lambda_7\lambda_8 L_{12} + \lambda_7\lambda_9 L_{13} + \lambda_7\lambda_{10} L_{14} + \lambda_8\lambda_9 L_{23} + \lambda_8\lambda_{10} L_{24} + \lambda_9\lambda_{10} L_{34})]/(N - 1),$$

$$\hat{\sigma}_{12} = (\lambda_1\lambda_2 L_1 + \lambda_1\lambda_3 L_{12})/(N - 1),$$

$$\hat{\sigma}_{13} = (\lambda_1\lambda_4 L_1 + \lambda_1\lambda_5 L_{12} + \lambda_1\lambda_6 L_{13})/(N - 1),$$

$$\hat{\sigma}_{14} = (\lambda_1\lambda_7 L_1 + \lambda_1\lambda_8 L_{12} + \lambda_1\lambda_9 L_{13} + \lambda_1\lambda_{10} L_{14})/(N - 1),$$

$$\hat{\sigma}_{23} = [\lambda_2\lambda_4 L_1 + \lambda_3\lambda_5 L_2 + (\lambda_2\lambda_5 + \lambda_3\lambda_4) L_{12} + \lambda_2\lambda_6 L_{13} + \lambda_3\lambda_6 L_{23}]/(N - 1),$$

$$\hat{\sigma}_{24} = [\lambda_2\lambda_7 L_1 + \lambda_3\lambda_8 L_2 + (\lambda_2\lambda_8 + \lambda_3\lambda_7) L_{12} + \lambda_2\lambda_9 L_{13} + \lambda_2\lambda_{10} L_{14} + \lambda_3\lambda_9 L_{23} + \lambda_3\lambda_{10} L_{24}]/(N - 1),$$

and

$$\hat{\sigma}_{34} = [\lambda_4\lambda_7 L_1 + \lambda_5\lambda_8 L_2 + \lambda_6\lambda_9 L_3 + (\lambda_4\lambda_8 + \lambda_5\lambda_7) L_{12} + (\lambda_4\lambda_9 + \lambda_6\lambda_7) L_{13} + (\lambda_5\lambda_9 + \lambda_6\lambda_8) L_{23} + \lambda_4\lambda_{10} L_{14} + \lambda_5\lambda_{10} L_{24} + \lambda_6\lambda_{10} L_{34}]/(N - 1).$$

These sample estimates may be used to construct estimates of genetic parameters which have known population values. The heritability of trait I may be estimated by

$$h_I^2 = 2\hat{\sigma}_{13}/\hat{\sigma}_1\hat{\sigma}_3$$

and of trait II by

$$h_{II}^2 = 2\hat{\sigma}_{24}/\hat{\sigma}_2\hat{\sigma}_4.$$

The genetic correlation is estimated from residual covariance components according to Hazel [1943] by four methods:

$$g_1 = \hat{\sigma}_{23}/(\hat{\sigma}_{13}\hat{\sigma}_{24})^{1/2}, \qquad g_2 = \hat{\sigma}_{14}/(\hat{\sigma}_{13}\hat{\sigma}_{24})^{1/2},$$

$$g_3 = (\hat{\sigma}_{14} + \hat{\sigma}_{23})/2(\hat{\sigma}_{13}\hat{\sigma}_{24})^{1/2}, \quad \text{and} \quad g_4 = (\hat{\sigma}_{14}\hat{\sigma}_{23}/\hat{\sigma}_{13}\hat{\sigma}_{24})^{1/2}.$$

It should be noted that the genetic model described here assumes a random mating population where gene effects approximate to the four-variate normal distribution. And, of course, if selection is applied to the population then the parameters will probably change. This model also does not consider effects due to linkage, sexual or cytoplasmic differences.

SAMPLING RESULTS

Samples were generated with $N = 100$ for the four-variate sampling model which has been described. Twelve hundred sets of the L's were obtained with 99 degrees of freedom for each set. These 1200 sampling coefficients were then combined to form 120 sets of L's with 990 degrees of freedom and 240 sets with 495 degrees of freedom. Then 24 sets of parameter values (see Table 2) were used with the sampling coefficients to construct sample estimates of the genetic correlations (g_1, g_2, g_3, and g_4). The true genetic correlation, g_0, and the heritabilities of the two traits, h_I^2 and h_{II}^2, for each parameter set are shown in Table 2.

The parameter values were assigned as follows: nine parameter sets (1–3, 7–9, and 13–15) are constructed so that the two traits have equal variances and heritabilities; nine other parameter sets (16–24) are constructed so that the variances of trait II are half the variances of trait I and the heritabilities of the two traits are equal; and the remaining six parameter sets (4–6 and 10–12) are constructed so that the two traits have equal variances but unequal heritabilities. Three values were chosen for heritabilities and genetic correlations, 0.2, 0.5, and 0.8 with the exception of sets 19 and 20 where the genetic correlations were .14 and .35.

The genetic interpretation of parameter sets 13, 15, and 22 is not possible. Inadvertently, covariances used in these sets were assigned which do not admit explanation by the usual genetic theory in that the environmental covariances between traits I and II exceed the environmental variances of traits I and II. Another inconsistency occurs in the parameter sets 4–6 and 10–12 where the phenotypic correlations between traits I and II are different for parent and offspring thus implying different environmental correlations. These differences were not intended but they do not invalidate comparisons with the theoretical variances since the derived estimates of the sampling variances depend only on a four variate distribution not necessarily subject to genetic interpretation. These inconsistencies should, however, be noted.

The means of the sample estimates are presented in Table 3. Estimates were discarded if the denominator of the estimate had a negative component or if the signs of the numerator components of g_4 were different. If the signs of the numerator components of g_4 were both negative, the sign of g_4 was considered to be negative.

TABLE 2

Parameter Values Used in Constructing Estimates of Genetic Correlations

Parameter Set	h_I[a]	h_{II}[b]	g_0[c]	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_{12}	σ_{13}	σ_{14}	σ_{23}	σ_{24}	σ_{34}
1	.20	.20	.20	1000	1000	1000	1000	100	100	20	20	100	100
2	.20	.20	.50	1000	1000	1000	1000	100	100	50	50	100	100
3	.20	.20	.80	1000	1000	1000	1000	100	100	80	80	100	100
4	.50	.20	.20	1000	1000	1000	1000	250	250	32	32	100	100
5	.50	.20	.50	1000	1000	1000	1000	250	250	79	79	100	100
6	.50	.20	.80	1000	1000	1000	1000	250	250	126	126	100	100
7	.50	.50	.20	1000	1000	1000	1000	250	250	50	50	250	250
8	.50	.50	.50	1000	1000	1000	1000	250	250	125	125	250	250
9	.50	.50	.80	1000	1000	1000	1000	250	250	200	200	250	250
10	.50	.80	.20	1000	1000	1000	1000	250	250	63	63	400	400
11	.50	.80	.50	1000	1000	1000	1000	250	250	158	158	400	400
12	.50	.80	.80	1000	1000	1000	1000	250	250	253	253	400	400
13	.80	.80	.20	1000	1000	1000	1000	400	400	80	80	400	400
14	.80	.80	.50	1000	1000	1000	1000	400	400	200	200	400	400
15	.80	.80	.80	1000	1000	1000	1000	400	400	320	320	400	400
16	.20	.20	.20	1000	500	1000	500	100	100	14	14	50	100
17	.20	.20	.50	1000	500	1000	500	100	100	35	35	50	100
18	.20	.20	.80	1000	500	1000	500	100	100	57	57	50	100
19	.50	.50	.14	1000	500	1000	500	250	250	25	25	125	250
20	.50	.50	.35	1000	500	1000	500	250	250	62	62	125	250
21	.50	.50	.50	1000	500	1000	500	250	250	88	88	125	250
22	.80	.80	.20	1000	500	1000	500	400	400	57	57	200	400
23	.80	.80	.50	1000	500	1000	500	400	400	141	141	200	400
24	.80	.80	.80	1000	500	1000	500	400	400	226	226	200	400

[a] Heritability of trait I. [b] Heritability of trait II. [c] Genetic correlation between traits I and II.

All four methods of estimation appear to provide unbiased estimates of the genetic correlation when the sample size is large. For small sample size ($N = 100$) and low heritability (.20) of at least one trait, however, the means presented in Table 3 indicate the estimates may be biased upwards. The bias apparently increases with an increase in the genetic correlation when heritability is fixed. It is also worth noting that the bias is larger in nearly every case for g_4 than for g_1 , g_2 or g_3

TABLE 3

Means of Sample Estimates of Genetic Correlations—Associated with Parameters in Table 2.[a]

Parameter Set	1200 Samples $f = 99$				240 Samples $f = 495$				120 Samples $f = 990$			
	g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4	g_1	g_2	g_3	g_4
1	.22	.25	.23	.31	.22	.17	.19	.25	.20	.16	.18	.20
2	.56	.65	.61	.78	.56	.51	.54	.57	.50	.46	.48	.50
3	1.06	1.12	1.09	1.24	.95	.89	.92	.89	.84	.79	.82	.78
4	.21	.15	.18	.21	.22	.18	.20	.23	.22	.17	.20	.23
5	.56	.51	.53	.58	.52	.50	.51	.49	.51	.48	.49	.48
6	.95	.89	.92	.95	.84	.81	.83	.80	.82	.78	.80	.78
7	.18	.15	.17	.21	.20	.18	.19	.20	.20	.18	.19	.19
8	.53	.50	.52	.53	.50	.48	.49	.48	.50	.48	.49	.48
9	.88	.86	.87	.83	.80	.79	.79	.78	.79	.78	.79	.78
10	.18	.15	.17	.21	.19	.18	.19	.19	.19	.18	.19	.18
11	.52	.50	.51	.50	.50	.49	.50	.49	.50	.49	.49	.49
12	.86	.83	.85	.81	.80	.79	.80	.79	.80	.79	.79	.79
13	.18	.18	.18	.21	.20	.19	.19	.19	.20	.19	.19	.19
14	.50	.49	.49	.48	.50	.49	.49	.49	.50	.49	.49	.49
15	.82	.80	.81	.79	.80	.79	.80	.79	.80	.79	.79	.79
16	.22	.15	.18	.12	.18	.14	.16	.21	.17	.14	.16	.17
17	.59	.61	.60	.70	.56	.52	.54	.56	.50	.47	.49	.50
18	1.09	.97	1.03	1.15	.94	.90	.92	.89	.84	.80	.82	.79
19	.10	.06	.08	.10	.14	.11	.12	.15	.14	.11	.13	.15
20	.34	.30	.32	.37	.35	.33	.34	.33	.35	.33	.34	.33
21	.52	.47	.49	.51	.50	.47	.48	.47	.50	.47	.48	.48
22	.17	.17	.17	.20	.19	.19	.19	.19	.20	.20	.20	.19
23	.49	.47	.48	.48	.50	.49	.49	.49	.50	.49	.49	.49
24	.80	.79	.80	.78	.80	.79	.79	.79	.80	.79	.79	.79

[a] f is the number of degrees of freedom associated with each sample.

when there is evidence of bias. This increased bias is probably caused by discarding estimates of g_4 when the numerator covariances differ in sign since in these cases g_1, g_2 or g_3 would usually be small or negative.

The sampling variances of the estimates of genetic correlations were computed as $\sum_{i=1}^{m_j} (g_{ji} - \bar{g}_{j.})^2/(m_j - 1)$ where m_j is the number of

TABLE 4

SAMPLING VARIANCES OF ESTIMATES OF GENETIC CORRELATIONS—ASSOCIATED WITH PARAMETERS IN TABLE 2.

Parameter Set	1200 Samples $f = 99$						240 Samples $f = 495$						120 Samples $f = 990$					
	$\hat{V}(g_1)$	$\hat{V}(g_2)$	$\hat{V}(g_3)$	*d.f.*[a]	$\hat{V}(g_4)$	*d.f.*[b]	$\hat{V}(g_1)$	$\hat{V}(g_2)$	$\hat{V}(g_3)$	*d.f.*[a]	$\hat{V}(g_4)$	*d.f.*[b]	$\hat{V}(g_1)$	$\hat{V}(g_2)$	$\hat{V}(g_3)$	*d.f.*[a]	$\hat{V}(g_4)$	*d.f.*[b]
1	9.148	4.749	3.202	842	4.365	436	.325	.339	.163	237	.196	121	.112	.131	.061	119	.055	78
2	8.757	7.444	2.580	845	3.204	492	.428	.376	.229	237	.199	179	.126	.132	.068	119	.049	119
3	8.440	7.678	4.110	849	4.313	569	.660	.519	.411	237	.381	222	.154	.145	.088	119	.086	117
4	.899	1 039	.436	1013	.459	563	.087	.100	.047	238	.041	153	.042	.048	.023	119	.015	87
5	.940	1.180	.500	1012	.436	695	.100	.101	.055	238	.018	224	.051	.045	.026	119	.026	117
6	1 344	1.080	.717	1003	.631	805	.138	.130	.085	238	.086	237	.071	.051	.037	119	.036	119
7	.231	.289	.127	1187	.129	722	.027	.031	.015	239	.012	185	.013	.017	.008	119	.006	103
8	.219	.267	.121	1188	.096	962	.027	.030	.014	239	.015	237	.013	.016	.008	119	.008	119
9	.315	.281	.161	1187	.145	1133	.030	.032	.017	239	.017	239	.015	.016	.008	119	.008	119
10	.130	.162	.069	1192	.066	759	.016	.018	.009	239	.008	210	.008	.009	.005	119	.005	114
11	.124	.120	.057	1192	.050	1059	.016	.016	.008	239	.008	239	.007	.009	.004	119	.004	119
12	.229	.123	.089	1192	.074	1177	.018	.017	.009	239	.009	239	.008	.009	.004	119	.004	119
13	.060	.067	.038	1199	.034	843	.010	.010	.006	226	.005	239	.005	.006	.003	119	.003	118
14	.051	.057	.027	1199	.026	1136	.008	.009	.005	239	.005	239	.004	.005	.002	119	.003	119
15	.054	.056	.026	1199	.028	1199	.009	.010	.004	239	.004	239	.004	.005	.002	119	.002	119
16	6.559	9.550	7.264	848	5.904	443	.315	.341	.162	237	.202	120	.110	.129	.060	119	.056	77
17	5.020	6.840	3.789	848	2.404	504	.401	.355	.209	237	.180	182	.121	.127	.064	119	.047	100
18	12 894	4.134	4.546	846	3.221	574	.618	.481	.377	237	.341	222	.147	.138	.082	119	.080	117
19	.230	.302	.150	1187	.170	699	.027	.031	.016	239	.014	159	.013	.016	.008	119	.005	86
20	.219	.284	.116	1187	.094	844	.025	.028	.013	239	.011	226	.012	.015	.007	119	.008	119
21	.276	.316	.103	1188	.080	968	.024	.027	.012	239	.013	237	.012	.014	.007	119	.007	119
22	.059	.067	.043	1199	.039	901	.010	.010	.007	239	.006	227	.004	.005	.003	119	.003	119
23	.041	.049	.024	1199	.021	1137	.007	.008	.004	239	.004	239	.003	.004	.002	119	.002	119
24	.037	.040	.015	1199	.017	1199	.006	.007	.003	239	.003	239	.003	.004	.001	119	.001	119

[a] Degrees of freedom of sampling variances of estimates of g_1, g_2 and g_3.

[b] Degrees of freedom of sampling variances of estimates of g_4.

estimates for the jth estimation method. These sampling variances are shown in Table 4.

Reeve [1955] gives approximate formulae for the large sample variances of the estimates g_3 and g_4. The variances are equal for the two estimates. This variance is (in our notation):

$$V(g) = \frac{g^2}{4f}\left(\frac{\sigma_1^2\sigma_4^2}{\sigma_{14}^2} + \frac{\sigma_2^2\sigma_3^2}{\sigma_{23}^2} + \frac{\sigma_1^2\sigma_3^2}{\sigma_{13}^2} + \frac{\sigma_2^2\sigma_4^2}{\sigma_{24}^2} + \frac{2\sigma_{12}\sigma_{34}}{\sigma_{14}\sigma_{23}} + \frac{2\sigma_{13}\sigma_{24}}{\sigma_{14}\sigma_{23}} - \frac{2\sigma_1^2\sigma_{34}}{\sigma_{14}\sigma_{13}}\right.$$
$$\left. - \frac{2\sigma_{12}\sigma_4^2}{\sigma_{14}\sigma_{24}} - \frac{2\sigma_{12}\sigma_3^2}{\sigma_{23}\sigma_{13}} - \frac{2\sigma_2^2\sigma_{34}}{\sigma_{23}\sigma_{24}} + \frac{2\sigma_{12}\sigma_{34}}{\sigma_{13}\sigma_{24}} + \frac{2\sigma_{14}\sigma_{23}}{\sigma_{13}\sigma_{24}} - 4\right);$$

where f is the degrees of freedom associated with each component estimated. Using the same method as Reeve the variances of g_1 and g_2 can be derived as

$$V(g_1) = \frac{g_1^2}{f}\left(\frac{\sigma_2^2\sigma_3^2}{\sigma_{23}^2} + \frac{\sigma_1^2\sigma_3^2}{4\sigma_{13}^2} + \frac{\sigma_2^2\sigma_4^2}{4\sigma_{24}^2} - \frac{\sigma_{12}\sigma_3^2}{\sigma_{23}\sigma_{13}} - \frac{\sigma_2^2\sigma_{34}}{\sigma_{23}\sigma_{24}} + \frac{\sigma_{12}\sigma_{34} + \sigma_{14}\sigma_{23}}{2\sigma_{13}\sigma_{24}} - \frac{1}{2}\right)$$

and

$$V(g_2) = \frac{g_2^2}{f}\left(\frac{\sigma_1^2\sigma_4^2}{\sigma_{14}^2} + \frac{\sigma_1^2\sigma_3^2}{4\sigma_{13}^2} + \frac{\sigma_2^2\sigma_4^2}{4\sigma_{24}^2} - \frac{\sigma_1^2\sigma_{34}}{\sigma_{14}\sigma_{13}} - \frac{\sigma_{12}\sigma_4^2}{\sigma_{14}\sigma_{24}} + \frac{\sigma_{12}\sigma_{34} + \sigma_{14}\sigma_{23}}{2\sigma_{13}\sigma_{24}} - \frac{1}{2}\right).$$

The large sample variances of the genetic correlations corresponding to the 24 parameter sets are listed in Table 5. The expected variances of g_1 and g_2 are about double those of g_3 and g_4. The expected variances of g_1 and g_2 are slightly different for parameter sets 4–6 and 10–12 for which the heritabilities of the two traits are different.

The computed sampling variances of the estimates of g_1 and g_2 are approximately twice those of g_3 and g_4 as expected. The agreement between the expected and computed sampling variances of the estimates is poor for the small sample size ($f = 99$) except for the cases in which the heritability of both traits is high (0.80). The differences between computed and expected variances are much less for $f = 495$. The variances of estimates which still deviate most from the expected are those associated with low heritability (0.20) of both traits. For a relatively large sample size ($f = 990$) the agreement between expected and computed values is surprisingly close for all combinations of parameters used in this study.

The extremely large sampling variances which occurred with some parameter sets are probably due to large estimates in turn due to small values of one or both of the denominator covariances. This suggests that heritability plays a dominant role in determining the sampling variances of genetic correlation estimates. The coefficient of variation for one of the denominator covariances under the normal distribution

TABLE 5

Expected Variances of Genetic Correlations for the Parameters of Table 2.[a]

Parameter Set	$f = 99$			$f = 495$			$f = 990$		
	$V(g_1)$	$V(g_2)$	$V(g_3)$	$V(g_1)$	$V(g_2)$	$V(g_3)$	$V(g_1)$	$V(g_2)$	$V(g_3)$
1	.990	.990	.495	.198	.198	.099	.099	.099	.050
2	1.036	1.036	.540	.207	.207	.108	.104	.104	.054
3	1.174	1.174	.676	.235	.235	.135	.117	.117	.068
4	.390	.379	.192	.078	.076	.038	.039	.038	.019
5	.414	.385	.206	.083	.077	.041	.041	.038	.021
6	.490	.444	.272	.098	.089	.054	.049	.044	.027
7	.148	.148	.078	.030	.030	.016	.015	.015	.008
8	.142	.142	.070	.028	.028	.014	.014	.014	.007
9	.151	.151	.077	.030	.030	.015	.015	.015	.008
10	.090	.089	.049	.018	.018	.010	.009	.009	.005
11	.083	.080	.040	.017	.016	.008	.008	.008	.004
12	.088	.082	.041	.018	.016	.008	.009	.008	.004
13	.054	.054	.033	.011	.011	.007	.005	.005	.003
14	.046	.046	.023	.009	.009	.005	.005	.005	.002
15	.045	.045	.020	.009	.009	.004	.004	.004	.002
16	.974	.974	.483	.195	.195	.097	.097	.097	.048
17	.994	.994	.503	.199	.199	.101	.099	.099	.050
18	1.113	1.113	.620	.223	.223	.124	.111	.111	.062
19	.147	.147	.081	.029	.029	.016	.015	.015	.008
20	.132	.132	.066	.026	.026	.013	.013	.013	.007
21	.126	.126	.059	.025	.025	.012	.013	.013	.006
22	.050	.050	.034	.010	.010	.007	.005	.005	.003
23	.037	.037	.019	.007	.007	.004	.004	.004	.002
24	.032	.032	.012	.006	.006	.002	.003	.003	.001

[a] The expected variance of g_4 is the same as that of g_3.

is $100\sqrt{(4 + h^4)/fh^4}$. Similarly we can define the average coefficient of variation for the two denominator covariances associated with the genetic correlation to be

$$A = \frac{100}{2\sqrt{f}}\left(\sqrt{\frac{4 + h_I^4}{h_I^4}} + \sqrt{\frac{4 + h_{II}^4}{h_{II}^4}}\right).$$

This average coefficient of variation turns out to be useful as a "rule of thumb" for determining when the theory approximation of the variance of the genetic correlation is likely to be good. We can note that when the average coefficient of variation, A, is 20 percent or less the theory approximations in Table 5 are quite closely in agreement with the empirical sampling variances summarized in Table 4.

This guide line would apply equally as well to estimates of genetic correlations as to their sampling variances. When the average coefficient of variation is larger than 20 percent then the estimates are likely to be biased upward. In fact, if prior knowledge indicates the average coefficient to be greater than 20 percent then some thought should be given to whether or not the genetic correlation should be estimated.

CONCLUSIONS

The results given in Table 4 indicate for sample sizes of 1,000 or more that the approximate formulae of Reeve [1955] for the large sample variance of the genetic correlation are accurate. For smaller sample sizes (500 or less) the approximations are not accurate unless the heritabilities of the examined traits are high. When the sample size is 100 or less the approximations may tend to be very misleading.

It seems safe to say that investigators who are estimating genetic correlations need at least 1,000 sets of observations in order to obtain reasonable estimates of the sampling variances of the estimates. Even then the sampling variances of the genetic correlation estimates may still be too large for the estimates to be of use especially if heritabilities of the traits are low, for example $h^2 \leq 0.20$. A useful guide may be the average coefficient of variation for the denominator covariances associated with the genetic correlation which depends on heritability. If this coefficient is 20 percent or less then the theory approximation may be quite good.

Estimation by either of procedures g_3 or g_4 seems preferable since the sampling variances are only about half as great as for methods g_1 or g_2. Procedure g_3 appears to be better than g_4 especially for small sample sizes since the results have indicated g_3 is less likely to be biased than g_4 probably because more estimates of g_4 will usually be discarded as imaginary or uninterpretable.

SUMMARY

A method of generating samples from a normally distributed population is described which has the advantage of being parameter free. Sample coefficients can be developed with zero means and unit variances and then different sets of parameter values used with the sample

coefficients to develop sample sums of squares and crossproducts from populations with different parameters. This method is used to construct sample estimates of genetic correlations for 3 sample sizes and 24 sets of parameter values. The sampling variances of these estimates are compared with those expected by application of Reeve's [1955] formulae.

ACKNOWLEDGMENTS

This work was supported by C. R. Henderson's National Science Foundation grant for the estimation of genetic parameters. The reviewer who suggested the use of the coefficient of variation of the denominator covariance as a possible guide to when the theoretical approximation may be quite good is thanked for his suggestion.

REFERENCES

Hazel, L. N. [1943]. The genetic basis for constructing selection indexes. *Genetics 28,* 476.

Mode, C. J., and Robinson, H. F. [1959]. Pleiotropism and the genetic variance and covariance. *Biometrics 15,* 518.

Reeve, E. C. [1955]. The variance of the genetic correlation coefficient. *Biometrics 11,* 357.

Robertson, A. [1959]. The sampling variance of the genetic correlation coefficient. *Biometrics 15,* 469.

Tallis, G. M. [1959]. Sampling errors of genetic correlation coefficients calculated from analyses of variance and covariance. *The Australian Jour. Stat. 1,* 35.

A NEW METHOD OF TESTING HYPOTHESES AND ESTIMATING PARAMETERS FOR THE LOGISTIC MODEL[1]

James E. Grizzle
Department of Biostatistics, School of Public Health
University of North Carolina, Chapel Hill, North Carolina, U.S.A.

1. INTRODUCTION

Data collected in many types of research often consist of the proportion of experimental units having a specified attribute. In this paper we shall be concerned with the analysis of this type of data when it can be arranged in a multiway classification as for example, a factorial arrangement. An example of this type of data is given in Table 1.

Several methods may be used in analyzing these data. Regardless of the method, a model relating the proportion having the attribute to the treatments must be assumed for a meaningful interpretation of the experimental results. The problem of the most appropriate model becomes particularly acute when some of the treatments are applied at several levels and it is desired to investigate the nature of treatment effects with regard to both main effects and interactions. If the sample sizes in the cells are equal, the observed proportions are often analyzed by the analysis of variance. A more desirable procedure is to analyze the arc sine transformation of the proportion. It is well known that this transformation stabilizes the variance if the sample sizes are equal and not too small, and, for a large class of data, it provides a unit of measurement on which treatment effects are approximately linear except at values of the proportion near zero or one.

In bioassay both the logit and the probit transformations have a long history of use. With appropriate extensions these models can be used in analyzing data of the type being discussed. Since the proportion responding has been found to increase sigmoidally with increasing stimulus for many phenomena, these transformations are particularly effective in providing a scale on which treatment effects are linear. Dyke and Patterson [1952] gave an example of how the logit trans-

[1] This research was sponsored by Cancer Chemotherapy National Service Center of the National Institutes of Health as part of a research contract.

formation can be used in analyzing a 2^n factorial arrangement. Their method can be adapted to other arrangements by an obvious extension.

The purpose of this paper is to present a new method of testing hypotheses when the logistic model is used. It is shown that estimates of the cell probabilities required to make certain tests can be used to obtain maximum likelihood (ML) estimates of the parameters in the assumed model with less computation than is commonly associated with ML estimation with models of this type. The approach used here was suggested by the work of Reiersøl [1954] and Mitra [1955].

2. NOTATION

To avoid multiple subscripts we will number the cells $i = 1, \cdots, r$. It will be assumed that the relationship between the treatments and the probability P_i of a response is described by

$$\mathbf{1} = [\log_e (P_i/Q_i)] = \mathbf{A}\boldsymbol{\theta}, \qquad k < r,$$

where $\mathbf{A}$ is an r by k non-singular matrix of known constants determined by the model and the design of the experiment, $\boldsymbol{\theta}$ is a k-element column vector of unknown parameters, $\mathbf{1}$ is a k-element column vector, and $Q_i = 1 - P_i$.

The following additional notation will be used:

n_i = sample size in the i-th cell,
u_i = number of responses observed,
$v_i = n_i - u_i$,
$\mathbf{C}_{i.}$ is the i-th row vector and
$\mathbf{C}_{.i}$ is the i-th column vector of a matrix $\mathbf{C}$.

3. THEORY

Mitra [1955] and Diamond [1958], proceeding along the same lines of proof as Cramér [1945], have given the mathematical properties that the model and the hypotheses must have for the test statistics to be asymptotically distributed as χ^2 when the null hypothesis is true, and for the existence of unique consistent estimates of the parameters in the model for samples from multinomial distributions. Their results are given as general forms with the model and the hypotheses being unspecified except for certain analytic conditions. It is easily demonstrated that the logistic model and tests of linear hypotheses have the required properties, Grizzle [1960].

3.1 *Tests of Hypotheses.*

Let it be given that the logistic model fits the data except for chance deviation. If we wish to test the hypothesis

$$H_0 : \mathbf{C}^*\boldsymbol{\theta} = \boldsymbol{\tau}, \tag{3.1.1}$$

where rank $\mathbf{C}^* = t, t \leq k < r$, and $\boldsymbol{\tau}$ is a vector of preassigned constants, we construct a matrix $\mathbf{C}$ such that

$$\mathbf{Cl} = \mathbf{C}^*\boldsymbol{\theta} = \boldsymbol{\tau}, \tag{3.1.2}$$

where $\mathbf{C}^* = \mathbf{CA}$, if H_0 is true. Then using (3.1.2) as restraints on the likelihood, estimates of P_i are obtained.

The logarithm of the likelihood subject to the restraints is

$$\phi = \text{constant} + \sum (u_i \log P_i + v_i \log Q_i) - \boldsymbol{\lambda}'(\mathbf{Cl} - \boldsymbol{\tau}), \tag{3.1.3}$$

where $\boldsymbol{\lambda}$ is a $t \times 1$ vector of Lagrangian multipliers.

Taking partial derivatives, equating them to zero and solving for $\hat{P}_i$, we find

$$\hat{P}_i = (u_i - \boldsymbol{\lambda}'\mathbf{C}_{.i})/n_i , \qquad i = 1, \cdots , r. \tag{3.1.4}$$

To complete the solution, $\boldsymbol{\lambda}$ must be eliminated from (3.1.4).

Let $\hat{\mathbf{l}}$ be the vector of elements $\hat{l}_i = \log_e (\hat{P}_i/\hat{Q}_i)$, where $\hat{P}_i$ is defined by (3.1.4). Then (3.1.2) implies

$$\mathbf{C}\hat{\mathbf{l}} = \boldsymbol{\tau}, \tag{3.1.5}$$

or its antilog

$$\prod_{i=1}^{r} \left(\frac{u_i - \boldsymbol{\lambda}'\mathbf{C}_{.i}}{v_i + \boldsymbol{\lambda}'\mathbf{C}_{.i}}\right)^{c_{ji}} = e^{\tau_j} , \qquad j = 1, \cdots , t, \tag{3.1.6}$$

can be solved for $\boldsymbol{\lambda}$. Either of these sets of equations can be solved by the familiar Newton-Raphson method which will be given in Section 4.1. In general, the form to use for ease of solution depends on c_{ij} .

4. COMPUTATION

4.1 *General Solution.*

Since it is not possible to derive an explicit solution for $\boldsymbol{\lambda}$ from (3.1.5) or (3.1.6), we resort to the Newton-Raphson method of solution. Thus we expand (3.1.5) about some guessed value of $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}_0$ say, neglect all derivatives higher than the first-order and solve the resulting equation for $\Delta\boldsymbol{\lambda} = \boldsymbol{\lambda} - \boldsymbol{\lambda}_0$ .Expanding $\mathbf{C}\hat{\mathbf{l}}$ about the point $\boldsymbol{\lambda}_0$, the first-order approximation to (3.1.5) is given by

$$\mathbf{Cl}_0 - \mathbf{C\,D}_0\mathbf{C}'\,\Delta\boldsymbol{\lambda} = \boldsymbol{\tau}, \tag{4.1.1}$$

where $\mathbf{D}$ is a diagonal matrix of elements $1/(n_i\hat{P}_i\hat{Q}_i)$, and $\mathbf{l}_0$ and $\mathbf{D}_0$ are $\hat{\mathbf{l}}$ and $\mathbf{D}$ evaluated at $\boldsymbol{\lambda}_0$. Therefore,

$$\Delta\boldsymbol{\lambda} = (\mathbf{C\,D}_0\mathbf{C}')^{-1}(\mathbf{Cl}_0 - \boldsymbol{\tau}). \tag{4.1.2}$$

If, as in many problems encountered in practice, $\boldsymbol{\tau} = \mathbf{O}$, then

$$\boldsymbol{\Delta\lambda} = (\mathbf{C}\ \mathbf{D}_0\mathbf{C}')^{-1}\mathbf{C}\mathbf{l}_0\ . \tag{4.1.3}$$

A more accurate solution can be computed by letting $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 + \boldsymbol{\Delta\lambda}$ become the $\boldsymbol{\lambda}_0$ for a second iteration. The process is repeated until $\boldsymbol{\Delta\lambda}$ is as small as desired. Then, if H_0 is true,

$$X^2 = \boldsymbol{\lambda}'\mathbf{C}\ \mathbf{D}_0\mathbf{C}'\boldsymbol{\lambda}, \tag{4.1.4}$$

where $\mathbf{D}_0$ is taken from the last iteration, approaches the χ^2-distribution with t degrees of freedom as the n_i get large.

There are two special cases for which computational formulas can be derived which do not involve matrix inversion. The first, the test with one degree of freedom, can be obtained in an obvious way from (4.1.2). The second is the test for homogeneity of a group of treatment effects. This is particularly important because many tests of interaction can be put into this form.

4.2 *Computation for the Test of Homogeneity.*

For the method to be useful we must be able to write H_0 as $\theta_1 = \theta_2 = \cdots = \theta_t$. In some problems this is easily done; for others the model may have to be reparameterized, and for some it is not possible. This hypothesis can also be written in the form $\theta_1 - \theta_t = \theta_2 - \theta_t = \cdots = \theta_{t-1} - \theta_t = 0$ so that $\mathbf{C}$ implied by the test is non-singular. The notation will be more clear if we renumber the l_i to become l_{ij} , $i = 1, \cdots, t$, $j = 1, \cdots, s$.

Given that the model fits the data, to test the hypothesis

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_t$$

there must exist c_{ij} such that

$$\begin{gathered} \sum_j c_{1j}l_{1j} = \theta_1 \\ \cdots \\ \sum_j c_{tj}l_{tj} = \theta_t\ . \end{gathered} \tag{4.2.1}$$

Hence,

$$\begin{gathered} \sum_j c_{1j}l_{1j} - \sum_j c_{tj}l_{tj} = 0, \\ \cdots \\ \sum_j c_{ij}l_{ij} - \sum_j c_{tj}l_{tj} = 0, \\ \cdots \\ \sum_j c_{t-1,j}l_{t-1,j} - \sum_j c_{tj}l_{tj} = 0, \end{gathered} \tag{4.2.2}$$

if H_0 is true. Using (4.2.2) as restraints on the likelihood in estimating the P_{ij}, we find

$$\hat{P}_{ij} = (u_{ij} - c_{ij}\lambda_i)/n_{ij}\ , \quad i = 1, \cdots, t-1,\ j = 1, \cdots, s, \tag{4.2.3}$$

and

$$\hat{P}_{tj} = \left(u_{tj} - c_{tj}\sum_{i=1}^{t-1}\lambda_i\right)\bigg/ n_{tj}\ , \qquad j = 1, \cdots, s.$$

If we let $\sum_{i=1}^{t-1} \lambda_i = -\lambda_t$ so that $\sum_{i=1}^{t} \lambda_i = 0$, and substitute $\hat{P}_{ij}$ into (4.2.1) we have

$$\sum_j c_{1j} \log\left(\frac{u_{1j} - c_{1j}\lambda_1}{v_{1j} + c_{1j}\lambda_1}\right) = \cdots = \sum_j c_{tj} \log\left(\frac{u_{tj} - c_{tj}\lambda_t}{v_{tj} + c_{tj}\lambda_t}\right). \tag{4.2.4}$$

Expanding (4.2.4) about trial values λ_{i0} of the λ_i, where $\sum \lambda_{i0} = 0$, we have

$$T_1 - \delta\lambda_1 S_1 = \cdots = T_t - \delta\lambda_t S_t\ , \tag{4.2.5}$$

where

$$T_i = \sum_{j=1}^{s} c_{ij} \log\left(\frac{u_{ij} - c_{ij}\lambda_{i0}}{v_{ij} + c_{ij}\lambda_{i0}}\right),$$

$$S_i = \sum_{j=1}^{s} c_{ij}^2 \left(\frac{1}{u_{ij} - c_{ij}\lambda_{i0}} + \frac{1}{v_{ij} + c_{ij}\lambda_{i0}}\right),$$

and

$$\delta\lambda_i = \lambda_i - \lambda_{i0}\ .$$

Choose the i-th and k-th members of (4.2.5) and solve for $\delta\lambda_k$. Then

$$\delta\lambda_k = T_k/S_k - T_i/S_k + \delta\lambda_i S_i/S_k\ , \qquad k = 1, \cdots, t.$$

Now $\sum_k \delta\lambda_k = 0$, and thus

$$H - T_i/S + \delta\lambda_i S_i/S = 0,$$

where

$$H = \sum_k T_k/S_k \quad \text{and} \quad 1/S = \sum_k 1/S_k\ .$$

Therefore

$$\delta\lambda_i = (T_i - HS)/S_i\ . \tag{4.2.6}$$

The process is continued as in (4.1.3) until $\delta\lambda_i$ is sufficiently small. Then

$$X^2 = \sum \lambda_i^2 S_i\ ,$$

where S_i, taken from the last iteration, is asymptotically distributed as χ^2 with $t - 1$ degrees of freedom if H_0 is true.

If we change (4.2.4) to a product by taking the antilog of the equations, it represents a generalization of Norton's [1945] extension of Bartlett's [1935] test. The solution to the product form of (4.2.4) is given by

$$\delta\lambda_i = (R_i - 1/Sh)/R_iS_i , \tag{4.2.7}$$

where

$$R_i = \prod_j \left(\frac{u_{ij} - c_{ij}\lambda_{i0}}{v_{ij} + c_{ij}\lambda_{i0}}\right)^{c_{ij}}, \qquad i = 1, \cdots ; t,$$

S_i and $1/S$ are as previously defined and $h = \sum_i 1/(R_iS_i)$. For small values of c_{ij}, (4.2.7) may be a more convenient computational form than (4.2.6) because tables do not have to be consulted. Also in the case of (4.1.3) with one degree of freedom the product form may be more convenient. The solution is

$$\delta\lambda = (R_1 - R_2e^{\tau})/(R_1S_1 + R_2S_2e^{\tau}), \tag{4.2.8}$$

where

$$R_1 = \prod_i (u_i - c_{1i}\lambda_0)^{c_{1i}}, \qquad R_2 = \prod_i (v_i + c_{1i}\lambda_0)^{c_{1i}},$$

$$S_1 = \sum_i c_{1i}^2/(u_i - c_{1i}\lambda_0) \quad \text{and} \quad S_2 = \sum_i c_{1i}^2/(v_i + c_{1i}\lambda_0).$$

5. TESTING OF THE FIT OF THE MODEL AND FITTING THE MODEL

Even though the significance level and the operating characteristics of the tests are not what they are presumed to be, most investigators test the agreement of the model with the data before proceeding to make tests on the parameters in the model. The procedures described in Sections 3 and 4 can be adapted for this purpose.

5.1 *Test of the Fit of the Model.*

The method can be made intuitively plausible by drawing an analogy with the analysis of variance. Recall that the residual or error sum of squares in the analysis of variance for a factorial experiment can be regarded as the sum of squares due to high-order interactions. Hence this sum of squares can be computed as the sum of squares associated with a set of contrasts. To obtain the proper test statistic choose **C** as the set of contrasts which would be interpreted as interaction among effects included in the model and interpreted as error, and then estimate the P, subject to the restraint $\mathbf{Cl} = \mathbf{O}$. This is equivalent to requiring that there be no deviation from the model, which is the hypothesis one wants to test. Once $\mathbf{C}$ is ascertained, the estimation and testing proceed as previously outlined. The λ computed in the process of making this test can be utilized to compute the ML estimate of $\boldsymbol{\theta}$ without the iteration required by conventional methods.

5.2 *Estimation of Parameters.*

To ascertain the relationship between λ and the observed logit, let us expand l_i about the point P_{i0}, the value of P_i if the data fit the model. Then, to a first-order approximation

$$\log_e (P_i/Q_i) = \log_e (P_{i0}/Q_{i0}) + (P_i - P_{i0})/(P_{i0}Q_{i0}).$$

Now $\lambda'\mathbf{C}_{.i} = P_i - P_{i0}$. Solving for $\log_e (P_{i0}/Q_{i0})$ and replacing P_i and Q_i by their estimates, u_i/n_i and v_i/n_i, we find

$$\log_e (\hat{P}_{i0}/\hat{Q}_{i0}) = \log_e (u_i/v_i) - (\lambda'\mathbf{C}_{.i})/(\hat{P}_{i0}\hat{Q}_{i0}),$$

where

$$\hat{P}_{i0} = (u_i - \lambda'\mathbf{C}_{.i})/n_i .$$

Therefore we see that $\log_e (\hat{P}_{i0}/\hat{Q}_{i0})$ taken from the test for the fit of the model represents the working logit.

Thus if we obtain estimates of the $\hat{P}_i$ subject to a suitably chosen set of restraints we can use them in the equation,

$$\mathbf{A'WA\theta} = \mathbf{A'Wl^*},$$

where $\mathbf{W}$ is a diagonal matrix of elements $n_i\hat{P}_{i0}\hat{Q}_{i0}$ and $\mathbf{l}^*$ has elements $\log_e (\hat{P}_{i0}/\hat{Q}_{i0})$, to estimate $\boldsymbol{\theta}$ without the iteration ordinarily associated with the solution of ML equations of this type.

Although iteration is not required to compute $\hat{\boldsymbol{\theta}}$ by this method, it is required for computing λ. By conventional methods a $k \times k$ matrix must be inverted for each iteration, but by this method a $(r - k) \times (r - k)$ matrix is inverted in each iteration in computing λ and a $k \times k$ matrix is inverted once in computing $\hat{\boldsymbol{\theta}}$. Since usually $r - k < k$ in problems of the type envisioned here, there is some saving in computation time through the use of this method. Furthermore for some experiments, λ can be obtained by methods given in Section 4.2 without inverting matrices. Or, as in the example that follows, we may need to compute λ in the process of making the tests.

5.3 *Example.*

Cochran [1954] gives the following data:

TABLE 1

Data on Number of Mothers with Previous Infant Losses

Birth Order		No. of Mothers with		Total
		Losses	No Losses	
2	Problems	20	82	102
	Controls	10	54	64
3–4	Problems	26	41	67
	Controls	16	30	46
5+	Problems	27	22	49
	Controls	14	23	37

It is desired to compare the mothers of Baltimore school children who have been referred by their teachers as presenting behavioral problems to mothers of a comparable group of control children. For each mother it is recorded whether she had suffered any infant losses previous to the child in the study.

The model assumed is

$$l_{ij} = \mu + \alpha_i + \beta_j , \qquad \sum \alpha_i = \sum \beta_j = 0,$$

where α_i is the effect of problem or control depending on whether $i = 1$ or 2, and β_j is the effect of the j-th birth order, $j = 1, 2, 3$.

First we will test the assumption that the model proposed fits the data. This may also be regarded as the test of the hypothesis of no $\alpha\beta$-interaction. If there is no interaction

$$l_{11} - l_{12} = l_{21} - l_{22} = l_{31} - l_{32} . \tag{5.3.1}$$

A convenient form of (5.3.1) to use as restraints in estimating the P_{ij} is

$$\begin{aligned} l_{11} - l_{12} - l_{31} + l_{32} &= 0, \\ l_{21} - l_{22} - l_{31} + l_{32} &= 0. \end{aligned} \tag{5.3.2}$$

The estimates are:

$$\hat{P}_{11} = (20 - \lambda_1)/102, \ \hat{P}_{21} = (26 - \lambda_2)/67, \ \hat{P}_{31} = (27 + \lambda_1 + \lambda_2)/49,$$

$$\hat{P}_{12} = (10 + \lambda_1)/64 \ , \ \hat{P}_{22} = (16 + \lambda_2)/46, \ \hat{P}_{32} = (14 - \lambda_1 - \lambda_2)/37.$$

To complete the solution, we will use the antilog form.

Let $\lambda_1 + \lambda_2 = -\lambda_3$, so that $\lambda_1 + \lambda_2 + \lambda_3 = 0$. Then the equations

$$\left(\frac{20 - \lambda_1}{82 + \lambda_1}\right)\left(\frac{54 - \lambda_1}{10 + \lambda_1}\right) = \left(\frac{26 - \lambda_2}{41 + \lambda_2}\right)\left(\frac{30 - \lambda_2}{16 + \lambda_2}\right) = \left(\frac{27 - \lambda_3}{22 + \lambda_3}\right)\left(\frac{23 - \lambda_3}{14 + \lambda_3}\right),$$

must be solved for λ_1, λ_2, λ_3 subject to the restriction $\lambda_1 + \lambda_2 + \lambda_3 = 0$.

The starting values of λ_{i0} are determined as follows: The objective of the iteration can be regarded as to make $R_1 = R_2 = R_3$, since when this occurs all $\delta\lambda_i = 0$. Therefore for the first trial values only, R_1, R_2 and R_3 need to be computed. The initial values of $\lambda_{10} = \lambda_{20} = \lambda_{30} = 0$ can be used if there are no zeros in the success and failure classifications. If for this test on this set of data we start with all $\lambda_{i0} = 0$, five iterations are required for accuracy comparable to the solution given. Needless to say, some practice in choosing λ_{i0} can save computing time.

After some preliminary examination of the type described, the values of

$$\lambda_{10} = -.5, \qquad \lambda_{20} = -1.0, \qquad \lambda_{30} = 1.5$$

were chosen for use in the complete cycles of iteration. After the computation shown in Table 2 we find

$$\lambda_1 = -.503, \qquad \lambda_2 = -1.213, \qquad \lambda_3 = 1.716.$$

TABLE 2

Computation for Test of Interaction

	Iteration 1	Iteration 2
R_1	1.443009	1.443009
R_2	1.395000	1.443979
R_3	1.505148	1.444384
S_1	.184662	.184662
S_2	.160962	.161549
S_3	.192797	.192333
$1/(R_1S_1)$	3.752782	3.752782
$1/(R_2S_2)$	4.453510	4.286823
$1/(R_3S_3)$	3.446054	3.599673
$1/Sh$	1.443038	1.443789
$\delta\lambda_1$	.000	− .003
$\delta\lambda_2$	− .214	.001
$\delta\lambda_3$	.214	.002

Of course this is only the approximate solution. Computation was stopped at the second interation because $\delta\lambda_i$ were trivial, indicating that further iteration would have very little effect on the test statistic.

Then

$$\begin{aligned} X^2 &= \lambda_1^2 S_1 + \lambda_2^2 S_2 + \lambda_3^2 S_3 \\ &= .851 \end{aligned}$$

For a χ^2-test with two degrees of freedom and .05-significance level, the critical region is given by $X^2 > 5.991$. Therefore we do not reject the hypothesis of no interaction.

Now that we have some assurance that the model fits the data, we can proceed to test hypotheses about the parameters in the model. The objective of the study is to ascertain whether there is a difference between problems and controls. This may be stated as

$$H_0 : \alpha_1 - \alpha_2 = 0.$$

The restraint associated with this test is

$$l_{11} + l_{21} + l_{31} - l_{12} - l_{22} - l_{32} = 0,$$

which gives the equation

$$\left(\frac{20 - \lambda}{82 + \lambda}\right)\left(\frac{26 - \lambda}{41 + \lambda}\right)\left(\frac{27 - \lambda}{22 + \lambda}\right) = \left(\frac{10 + \lambda}{54 - \lambda}\right)\left(\frac{16 + \lambda}{30 - \lambda}\right)\left(\frac{14 + \lambda}{23 - \lambda}\right).$$

As a preliminary estimate of λ, choose $\lambda_0 = 2$. After two iterations we find $\lambda = 2.188$ and $X^2 = 2.475$. Therefore if we put the significance level at .05 we do not reject H_0. It is interesting to note that the probability of observing an $X^2 > 2.475$ is approximately 12 percent if H_0 is true while the test Cochran [1954] suggests has probability of approximately 10 percent.

Another hypothesis of interest is

$$H_0 : \beta_1 = \beta_2 = \beta_3 ,$$

that is, equality of birth-order effects. To test this hypothesis we use the restraints

$$l_{11} + l_{12} = l_{21} + l_{22} = l_{31} + l_{32} ,$$

which can be written

$$\begin{aligned} l_{11} + l_{12} - l_{31} - l_{32} &= 0, \\ l_{21} + l_{22} - l_{31} - l_{32} &= 0. \end{aligned}$$

Using the same technique as in the test for interaction, we obtain the equations

$$\left(\frac{20 - \lambda_1}{82 + \lambda_1}\right)\left(\frac{10 - \lambda_1}{54 + \lambda_1}\right) = \left(\frac{26 - \lambda_2}{41 + \lambda_2}\right)\left(\frac{16 - \lambda_2}{30 + \lambda_2}\right) = \left(\frac{27 - \lambda_3}{22 + \lambda_3}\right)\left(\frac{14 - \lambda_3}{23 + \lambda_3}\right).$$

Starting with $\lambda_{10} = -9$, $\lambda_{20} = 5$, $\lambda_{30} = 4$, after three iterations we find that

$$\lambda_1 = -9.925, \qquad \lambda_2 = 3.529, \qquad \lambda_3 = 6.396,$$

and X^2 with two degrees of freedom is 24.239. Therefore we reject H_0.

Now that there is evidence that birth order has a significant effect it will be instructive, as an example of the technique for testing a single contrast, to ascertain the nature of the effect. For simplicity let us assume that the birth orders 2, 3 – 4, 5+ are approximately equally spaced. Two hypotheses of interest are: is the effect of birth order linear or does it require a second-degree polynomial equation to describe the effect?

This can be ascertained by testing

$$H_{01} : \beta_1 - \beta_3 = 0,$$

and

$$H_{02} : \beta_1 - 2\beta_2 + \beta_3 = 0.$$

The restrictions for estimating the P_{ij} are

$$l_{11} + l_{12} - l_{31} - l_{32} = 0$$

and

$$l_{11} + l_{12} - 2l_{21} - 2l_{22} + l_{31} + l_{32} = 0$$

for testing H_{01} and H_{02} respectively. For the test of H_{01}, the equation is

$$\left(\frac{20 - \lambda}{82 + \lambda}\right)\left(\frac{10 - \lambda}{54 + \lambda}\right) = \left(\frac{27 + \lambda}{22 - \lambda}\right)\left(\frac{14 + \lambda}{23 - \lambda}\right).$$

Starting with $\lambda_0 = -7$ in three iterations we find λ to be -7.535.

The computations as given by (4.2.8) are shown in the following table.

TABLE 3

Computation for Test of a Single Contrast

	Iteration 1	Iteration 2	Iteration 3
R_1	.130213	.139294	.139544
R_2	.160919	.140070	.139537
S_1	.130470	.128352	.128297
S_2	.260673	.272320	.272661
$\delta\lambda$	−.521	−.014	.000

$$X^2 = \lambda^2(S_1 + S_2) = 22.765.$$

This is far beyond the .05-point for χ^2 with one degree of freedom. Therefore we reject H_{01}.

For H_{02} the equation is

$$\left(\frac{20 - \lambda}{82 + \lambda}\right)\left(\frac{10 - \lambda}{54 + \lambda}\right)\left(\frac{27 - \lambda}{22 + \lambda}\right)\left(\frac{14 - \lambda}{23 + \lambda}\right) = \left(\frac{26 + 2\lambda}{41 - 2\lambda}\right)^2\left(\frac{16 + 2\lambda}{30 - 2\lambda}\right)^2.$$

Starting with $\lambda_0 = -1$, after two iterations we find λ to be -1.191 and $X^2 = 1.478$. Therefore we do not reject H_{02} and we conclude that within the range covered by the data, birth-order effects are linear on the logistic scale.

For the data presented here for illustrative purposes, it is not particularly helpful to estimate the parameters in the model. However, we will proceed with the estimation to illustrate the non-iterative procedure for estimating parameters. The $\hat{P}_{ij}$ obtained under the no-interaction hypothesis are

$$\hat{P}_{11} = .2010, \quad \hat{P}_{21} = .4062, \quad \hat{P}_{31} = .5160,$$
$$\hat{P}_{12} = .1483, \quad \hat{P}_{22} = .3215, \quad \hat{P}_{32} = .4248,$$

and the associated l^*_{ij} are

$$l^*_{11} = -1.380, \quad l^*_{21} = -.380, \quad l^*_{31} = .064,$$
$$l^*_{12} = -1.750, \quad l^*_{22} = -.747, \quad l^*_{32} = -.302.$$

The model can be reparameterized in several ways to make the equations non-singular. One way is to let $\alpha = \alpha_2 - \alpha_1$, $\eta_1 = \beta_1 - \beta_2$, $\eta_2 = \beta_1 - \beta_3$, then

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \\ 1 & 1 & 0 & -1 \end{bmatrix},$$

and the normal equations are

$$\begin{bmatrix} 71.93772 & -17.62000 & -1.72980 & 3.18720 \\ -17.62000 & 71.93772 & -2.17140 & -5.10068 \\ -1.72980 & -2.17140 & 50.65988 & 24.46504 \\ 3.18720 & -5.10068 & 24.46504 & 45.74288 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} -52.3495 \\ 3.5944 \\ -23.1122 \\ -34.8062 \end{bmatrix}.$$

The solution is

$$\hat{\mu} = -.749, \qquad \hat{\alpha} = -.184, \qquad \hat{\eta}_1 = -.185, \qquad \hat{\eta}_2 = -.630,$$

and the variances of the estimates are .0149, .0149, .0267 and .0298 respectively. The predicted logits using these estimates are

$$l_{11}^* = -1.381, \qquad l_{21}^* = -.380, \qquad l_{31}^* = .064,$$
$$l_{12}^* = -1.748, \qquad l_{22}^* = -.748, \qquad l_{32}^* = -.303,$$

which are very close to those used as the original estimates. Unless more then three-place accuracy is desired no further iteration is necessary.

6. SUMMARY AND DISCUSSION

In this paper we have developed new methods for making tests on the parameters in the logistic model. For many types of data the methods developed here allow computation of X^2 for tests of hypotheses about some subset of the parameters in the model without having to estimate them all as is usually done when using the logistic model. This might be particularly useful in combining 2 × 2 tables, as in the test of problems versus controls in the example given, which can be regarded as three 2 × 2 tables. If all that we desire to do is to test the difference between problems and controls, we need to do only that part of the work done for testing $\alpha_1 - \alpha_2 = 0$ in the example, if we can assume that the logistic model fits the data.

Once λ associated with a properly chosen **C**-matrix has been determined, it can be used to compute the ML estimate of the parameters in the model without further iteration. The relationship between λ and the working logit and between the logistic model and Bartlett's test of interaction is shown.

ACKNOWLEDGMENTS

The author would like to express his appreciation to Dr. D. B. Duncan, who suggested a problem leading to this work, and to Dr. H. L. Lucas, Jr. for direction in preparing his Ph.D. dissertation, from which this paper was taken.

REFERENCES

Bartlett, M. S. [1935]. Contingency table interactions. *Supplement to the Jour. of the Royal Stat. Soc. 11*, 248–52.

Cochran, W. G. [1954]. Some methods for strengthening the common χ^2 tests. *Biometrics 10*, 417–51.

Cramér, H. [1945]. *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

Diamond, E. L. [1958]. *Asymptotic power and independence of certain classes of tests on categorical data.* North Carolina Institute of Statistics Mimeograph Series No. 196.

Dyke, G. V. and Patterson, H. D. [1952]. Analysis of factorial arrangements when the data are proportions. *Biometrics 8*, 1–12.

Grizzle, J. E. [1960]. *Application of the logistic model to analyzing categorical data.* North Carolina Institute of Statistics Mimeograph Series No. 251.

Lancaster, H. O. [1951]. Complex contingency tables treated by the partition of chi square. *Jour. of the Royal Stat. Soc., Series B, 13*, 242–9.

Mitra, S. K. [1955]. *Contributions to the statistical analysis of categorical data.* North Carolina Institute of Statistics Mimeograph Series No. 142.

Norton, H. W. [1945]. Calculation of chi squared for complex contingency tables. *Jour. of the Amer. Stat. Assoc. 40*, 251–8.

Reiersol, O. [1951]. Tests of linear hypotheses concerning binomial experiments. *Skandinavisk Aktuarietidskrift 37*, 38–59.

Roy, S. N. and Kastenbaum, M. A. [1956]. On the hypothesis of no "interaction" in a multiway contingency table. *Ann. of Math. Stat. 27*, 749–57.

A GENERALIZED MODEL OF A HOST-PATHOGEN SYSTEM

C. J. Mode
Department of Mathematics, Montana State College
Bozeman, Montana, U.S.A.

INTRODUCTION

This paper is a sequel to a paper, "A Model of a Host-pathogen System with Particular Reference to the Rusts of Cereals", which appeared in *Biometrical Genetics* [1960]. The former paper was restricted to the summer stage of the rusts of cereals, but in this paper a wider class of host-pathogen systems will be considered and some assumptions will be relaxed. Specifically, the assumptions of random association of the host and pathogen, equal number of varieties of the host and races of the pathogen, and constant fitness functions will be dropped.

The results of this paper are intended to apply to host-pathogen systems satisfying the following conditions:

1. The pathogen reproduces on the host.

2. The host may be differentiated into varieties on the basis of its resistance to the races of the pathogen.

3. The pathogen may be differentiated into races on the basis of its ability to grow on a set of host varieties.

4. Host resistance to a particular race of the pathogen is genetically controlled.

5. The damage to the host caused by the pathogen in a given time interval is directly related to the increase in number in the pathogen population during the given time interval.

It should be pointed out that conditions (1) and (4) imply that no assumptions are made with respect to the mode of reproduction of the pathogen and the mode of inheritance of host resistance to the pathogen.

Many economic crop plants and their foliar diseases caused by pathogenic fungi are examples of host-pathogen systems satisfying the above conditions. These host-pathogen systems are characterized by frequent shifts in the racial frequencies of the pathogen population, making it difficult to maintain host resistance to the pathogen. It seems plausible that damage to the host in such host-pathogen systems could be adequately controlled if (1) the racial structure (the relative frequencies of the races) in the pathogen population were stabilized and (2) the increases in number in the pathogen population was held below some critical number during the growing season of the crop.

In a previous paper, Mode [1960], the writer was unable to find conditions under which the racial structure of the pathogen population was stabilized. The purpose of this paper is two-fold, namely, (1) to characterize a host-pathogen system containing an arbitrary number of varieties of the host and an arbitrary number of races of the pathogen and (2) to find some conditions under which condition (1) of the previous paragraph is met. It will be assumed that the host-pathogen systems under consideration may be characterized in terms of continuous and differentiable functions of time.

In passing it is interesting to note that in at least one case, the genetics of both the host and the pathogen has been worked out. The reader is referred to the remarkable paper of Flor [1956] in which the complementary genetic systems of flax and flax rust are discussed. As an example of experimental work in the field under consideration, the reader is referred to the recent paper of Suneson [1960].

1. POPULATION NUMBERS AND THE ASSOCIATION OF THE HOST AND PATHOGEN

Let H be the number of members of the host population and let H_i be the number of the members belonging to variety $V_i (i = 1, \cdots, m)$ at time t. Similarly, let P be the number of members of the pathogen population and P_j the number of these members belonging to race $R_j (j = 1, \cdots, n)$ at time t. The relations of the H_i and P_j to H and P are given by

$$H = \sum_{i=1}^{m} H_i \quad \text{and} \quad P = \sum_{j=1}^{n} P_j \, . \tag{1.1}$$

Let a member h of the host population and a member p of the pathogen population be chosen at random. The probability that h belongs to variety V_i is

$$\Pr (h \, \varepsilon \, V_i) = H_i/H = x_i \, , \tag{1.2}$$

and the probability that p belongs to race R_j is

$$\Pr (p \, \varepsilon \, R_j) = P_j/P = y_j \, . \tag{1.3}$$

We shall consider next the association of the host and pathogen. Let φ_{ij} be the probability that at time t a member h of the host population belonging to variety V_i and a member p of the pathogen population belonging to race R_j are associated. In symbols we may write

$$\Pr\,(h\,\varepsilon\,V_i\,,\,p\,\varepsilon\,R_j) = \varphi_{ij}\,. \tag{1.4}$$

If the host and pathogen are associated at random, then we have independence in the probability sense so that

$$\varphi_{ij} = x_iy_i\,, \tag{1.5}$$

for all i and j.

In general, however, we would expect non-random association of the host and pathogen to be the rule due to the nature of the specific reaction of the host and pathogen to each other. In order to take nonrandomness into account we introduce a measure of departure from random association, θ_{ij} . The measure of departure from random association is a positive number satisfying the relation

$$\varphi_{ij} \doteq \theta_{ij}x_iy_j\,. \tag{1.6}$$

Kimura [1958] introduced this measure in connection with the study of nonrandom mating diploid populations. We note that the measure θ_{ij} is related to certain conditional probabilities. The conditional probability that a member p of the pathogen population belongs to race R_j given that it is associated with a member h of the host population belonging to variety V_i is

$$\Pr\,(p\,\varepsilon\,R_j \mid h\,\varepsilon\,V_i) = \varphi_{ij}/x_i = y_j\theta_{ij}\,. \tag{1.7}$$

Similarly, the conditional probability that a member h of the host population belongs to variety V_i given that it is associated with a member p of the pathogen population belonging to race R_j is

$$\Pr\,(h\,\varepsilon\,V_i \mid p\,\varepsilon\,R_j) = \varphi_{ij}/y_i = x_i\theta_{ij}\,. \tag{1.8}$$

Since $x_i\theta_{ij}$ and $y_j\theta_{ij}$ are conditional probabilities, it follows that

$$\sum_{i=1}^{m} x_i\theta_{ij} = \sum_{j=1}^{n} y_j\theta_{ij} = 1. \tag{1.9}$$

2. THE FITNESS FUNCTIONS

Due to the specific nature of the reaction of the host and pathogen to each other, with each association ij of the host and pathogen we shall associate two fitness functions, one for the host and one for the pathogen. The fitness functions may be regarded as measures of the ability of the host and pathogen to reproduce in a given association. Let λ_{ij} be the fitness function of the host and μ_{ij} the fitness function of the pathogen in the ij-th association. Since the ability of the host and pathogen to reproduce in a given association may depend on the

x_i and y_j, we shall allow λ_{ij} and μ_{ij} to be real-valued functions of the real variables x_i and y_j. Clearly, for every φ_{ij} we associate a λ_{ij} and μ_{ij} so that the pair (λ_{ij}, μ_{ij}) may be regarded as a random variable with respect to the φ_{ij}.

The expected value of λ_{ij} or the mean fitness of the host population is

$$E(\lambda) = \lambda_{..} = \sum_{ij} \varphi_{ij}\lambda_{ij} , \tag{2.1}$$

and the expected value of μ_{ij} or the mean fitness of the pathogen population is

$$E(\mu) = \mu_{..} = \sum_{ij} \varphi_{ij}\mu_{ij} . \tag{2.2}$$

The mean fitness of host variety V_i, or the conditional expectation of λ_{ij} given V_i, is

$$E(\lambda \mid V_i) = \lambda_{i.} = \sum_{j=1}^{n} y_j\theta_{ij}\lambda_{ij} , \tag{2.3}$$

and the mean fitness of race R_j of the pathogen, or the conditional expectation of μ_{ij} given R_j, is

$$E(\mu \mid R_j) = \mu_{.j} = \sum_{i=1}^{m} x_i\theta_{ij}\mu_{ij} . \tag{2.4}$$

The conditional expectations of λ_{ij} given R_j and μ_{ij} given V_i, which are also of interest, are given by

$$\begin{aligned} E(\lambda \mid R_j) &= \lambda_{.j} = \sum_{i=1}^{m} x_i\theta_{ij}\lambda_{ij} , \\ E(\mu \mid V_i) &= \mu_{i.} = \sum_{j=1}^{n} y_j\theta_{ij}\mu_{ij} . \end{aligned} \tag{2.5}$$

We now make the following definitions of the measures of fitness of variety V_i and race R_j. The measure of fitness of variety V_i of the host population is defined by the differential equation,

$$d(\log H_i)/dt = \lambda_{i.} , \tag{2.6}$$

and the measure of fitness of R_j of the pathogen population is defined by

$$d(\log P_j)/dt = \mu_{.j} . \tag{2.7}$$

These definitions of the measures of fitness of variety V_i and R_j are equivalent to those given in the previous paper.

From the definitions of the measure of fitness of variety V_i and race R_j, it may be shown that the change in the probabilities x_i and y_j

in time is characterized by the set of differential equations

$$\begin{aligned} dx_i/dt &= x_i(\lambda_{i.} - \lambda_{..}), \qquad (i = 1, \cdots, m), \\ dy_j/dt &= y_j(\mu_{.j} - \mu_{..}), \qquad (j = 1, \cdots, n). \end{aligned} \tag{2.8}$$

Finally, by differentiating H and P with respect to time, we may show that the changes in population number in the host and pathogen populations are given by

$$d(\log H)/dt = \lambda_{..}\,, \qquad d(\log P)/dt = \mu_{..}\,. \tag{2.9}$$

3. THE VARIATION AND COVARIATION IN FITNESS IN A HOST-PATHOGEN SYSTEM

We continue our characterization of the host-pathogen system by defining certain variances and covariances in fitness.

The total variance in fitness in the host population is

$$\text{var}\,(\lambda) = E(\lambda - \lambda_{..})^2 = \sum_{ij} \varphi_{ij}(\lambda_{ij} - \lambda_{..})^2, \tag{3.1}$$

the total variance in fitness in the pathogen population is

$$\text{var}\,(\mu) = E(\mu - \mu_{..})^2 = \sum \varphi_{ij}(\mu_{ij} - \mu_{..})^2, \tag{3.2}$$

and the total covariance in fitness in the host-pathogen system is

$$\text{cov}\,(\lambda, \mu) = E(\lambda - \lambda_{..})(\mu - \mu_{..}) = \sum_{ij} \varphi_{ij}(\lambda_{ij} - \lambda_{..})(\mu_{ij} - \mu_{..}). \tag{3.3}$$

In addition to the variance and covariance in fitness, we may also define certain components of variance and covariance which are useful in characterizing a host-pathogen system.

The variance in fitness in the host population attributable to varieties is

$$\text{var}\,(\lambda; V) = \sum_i x_i(\lambda_{i.} - \lambda_{..})^2, \tag{3.4}$$

the variance in fitness attributable to races is

$$\text{var}\,(\lambda; R) = \sum_j y_j(\lambda_{.j} - \lambda_{..})^2, \tag{3.5}$$

and the variance in fitness in the host population attributable to the interaction of varieties and races is

$$\text{var}\,(\lambda; VR) = \sum_{ij} \varphi_{ij}(\lambda_{ij} - \lambda_{i.} - \lambda_{.j} + \lambda_{..})^2, \tag{3.6}$$

By continuing in the same way we may define analogous components of variance in fitness for the pathogen population. Thus we shall let var (μ; V), var (μ; R), and var (μ; VR) stand for the components of variance in fitness in the pathogen population attributable to varieties, races, and the interaction of varieties and races, respectively. Similarly, we shall let cov $(\lambda, \mu; V)$, cov $(\lambda, \mu; R)$, and cov $(\lambda, \mu; VR)$ represent the components of covariance in fitness attributable to varieties, races, and the interaction of varieties and races respectively. These components of covariance are, of course, defined in the obvious way.

In particular, if the host and pathogen are associated at random so that $\theta_{ij} = 1$ for all i and j, then the following relations among the total variance and covariance and the components of variance and covariance hold.

$$\begin{aligned} \text{var}\ (\lambda) &= \text{var}\ (\lambda; V) + \text{var}\ (\lambda; R) + \text{var}\ (\lambda; VR), \\ \text{var}\ (\mu) &= \text{var}\ (\mu; V) + \text{var}\ (\mu; R) + \text{var}\ (\mu; VR), \qquad (3.7) \\ \text{cov}\ (\lambda, \mu) &= \text{cov}\ (\lambda, \mu; V) + \text{cov}\ (\lambda, \mu; R) + \text{cov}\ (\lambda, \mu; VR). \end{aligned}$$

Relations (3.7) will not hold in general, however, if the association of the host and pathogen is nonrandom.

4. THE CHANGE OF $\lambda_{..}$, $\mu_{..}$, VAR(λ), VAR(μ), AND COV(λ,μ) IN TIME

We complete our characterization on the host-pathogen system by finding differential equations characterizing the change in $\lambda_{..}$, $\mu_{..}$, var (λ), var (μ), and cov (λ, μ) in time. The proofs of the results of this section are easily obtained by straight-forward differentiation of the functions in question with respect to time and by using the results and definitions of the preceding sections.

When one wishes to find these differential equations, certain variables arise. A considerable simplification in representation may be gained if we set $\mathring{\theta}_{ij} = d(\log \theta_{ij})/dt$, $\mathring{\varphi}_{ij} = d(\log \varphi_{ij})/dt$, $\dot{\lambda}_{ij} = d\lambda_{ij}/dt$, and $\dot{\mu}_{ij} = d\mu_{ij}/dt$. The relations $\dot{\theta}_{ij} = \theta_{ij}\mathring{\theta}_{ij}$ and $\dot{\varphi}_{ij} = \varphi_{ij}\mathring{\varphi}_{ij}$ are also useful. Note, with each φ_{ij} we may associate the four-tuple, $(\mathring{\varphi}_{ij}$, $\mathring{\theta}_{ij}$, $\dot{\lambda}_{ij}$, $\dot{\mu}_{ij})$, so that the four-tuple may be regarded as a random variable with respect to φ_{ij} .

With the above conventions, the change in mean fitness in the host population becomes

$$d\lambda_{..}/dt = \text{var}\ (\lambda; V) + \text{cov}\ (\lambda, \mu; R) + E(\mathring{\theta}\lambda) + E(\dot{\lambda}), \qquad (4.1)$$

and the change in mean fitness in the pathogen population becomes

$$d\mu_{..}/dt = \text{var}\ (\mu; R) + \text{cov}\ (\lambda, \mu; V) + E(\mathring{\theta}\mu) + E(\dot{\mu}). \qquad (4.2)$$

Thus, the change in the mean fitness in time in the host population partitions into a variance component in the host population attributable to varieties, a component of covariance attributable to races, a term attributable to the changes in the measures of departure from random association, and a term attributable to changes in the fitness functions in time. The equation characterizing the change of mean fitness in time in the pathogen population is similar. If all θ_{ij} , λ_{ij} , and μ_{ij} are constant, then equations (4.1) and (4.2) reduce to the results given in the previous paper.

The differential equation characterizing the change in time of the total variance in fitness in the host population is

$$d[\text{var}\ (\lambda)]/dt = E[\mathring{\varphi}(\lambda - \lambda_{..})^2] + 2\ \text{cov}\ (\lambda, \dot{\lambda}), \tag{4.3}$$

and the differential equation giving the change in time of the total variance in fitness in the pathogen population is

$$d[\text{var}\ (\mu)]/dt = E[\mathring{\varphi}(\mu - \mu_{..})^2] + 2\ \text{cov}\ (\mu, \dot{\mu}). \tag{4.4}$$

Finally, the differential equation characterizing the change in the total covariance in fitness in time is

$$d[\text{cov}\ (\lambda, \mu)]/dt = E[\mathring{\varphi}(\lambda - \lambda_{..})(\mu - \mu_{..})] + \text{cov}\ (\dot{\lambda}, \mu) + \text{cov}\ (\lambda, \dot{\mu}). \tag{4.5}$$

It will be noted that in equations (4.1) through (4.5) we have set

$$E(\dot{\lambda}) = \sum_{ij} \varphi_{ij}\dot{\lambda}_{ij}\ , \qquad E(\mathring{\theta}\lambda) = \sum_{ij} \varphi_{ij}\mathring{\theta}_{ij}\lambda_{ij}\ ,$$
$$E[\mathring{\varphi}(\lambda - \lambda_{..})^2] = \sum_{ij} \varphi_{ij}\mathring{\varphi}_{ij}(\lambda_{ij} - \lambda_{..})^2. \tag{4.6}$$

$E(\mathring{\theta}\mu)$, $E(\dot{\mu})$, $E[(\mathring{\varphi}\mu - \mu_{..})^2]$, and $E[\mathring{\varphi}(\lambda - \lambda_{..})(\mu - \mu_{..})]$ are of course defined in a similar way. It is instructive to study the form of equations (4.3), (4.4), and (4.5) according as all θ_{ij} , λ_{ij} and μ_{ij} are constant or nonconstant.

5. STATIONARY STATE SYSTEMS

We shall say a host-pathogen system is in a stationary state if all φ_{ij} cease to change with time. Our discussion of stationary state systems begins by considering the case when all θ_{ij} , λ_{ij} , and μ_{ij} are constants. We note from equation (1.9) that, if all θ_{ij} are constant and the system is in a nonstationary state, then the x_i and y_j must change in such a way that the equation

$$\sum_{i,j} x_i y_j \theta_{ij} = 1 \tag{5.1}$$

is satisfied.

Let $\mathbf{A} = [a_{ij}] = [\theta_{ij}\lambda_{ij}]$ and $\mathbf{B} = [b_{ij}] = [\theta_{ij}\mu_{ij}]$, be constant $m \times n$ matrices, and let $\hat{\mathbf{x}}' = (\hat{x}_1, \cdots, \hat{x}_m)$ and $\hat{\mathbf{y}}' = (\hat{y}_1, \cdots, \hat{y}_n)$ be vectors of stationary state probabilities. In addition, let $\hat{\lambda}_{..}$ and $\hat{\mu}_{..}$ be the values of $\lambda_{..}$ and $\mu_{..}$ corresponding to the stationary state vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. Finally, let $\mathbf{c}_1$ and $\mathbf{c}_2$ represent $m \times 1$ and $n \times 1$ column vectors consisting of all $\hat{\lambda}_{..}$'s and $\hat{\mu}_{..}$'s respectively, i.e. $\mathbf{c}_1' = (\hat{\lambda}_{..}, \cdots, \hat{\lambda}_{..})$ and $\mathbf{c}_2' = (\hat{\mu}_{..}, \cdots, \hat{\mu}_{..})$. Throughout the remainder of the paper a prime will be used to denote the transpose of a matrix or vector.

From the definition of a stationary state system, it follows that four classes of stationary states may exist; namely, (1) when population number in both the host and pathogen populations is constant, (2) when population number in the host population is constant but that in the pathogen population is variable, (3) when population number in the host population is variable but that in the pathogen population is constant, and (4) when population number in both the host and pathogen population is variable. Henceforth we shall refer to the four classes of stationary state systems as systems of Class I, II, III, and IV, respectively.

If a host-pathogen system belongs to Class I, for example, then the defining equations of the class are

$$\begin{aligned} d(\log H)/dt &= \lambda_{..} = 0, \\ d(\log P)/dt &= \mu_{..} = 0. \end{aligned} \tag{5.2}$$

Moreover, if the system is in a stationary state, then the set of differential equations

$$\begin{aligned} dx_i/dt &= x_i\lambda_{i.} = 0 \qquad (i = 1, \cdots, m), \\ dy_j/dt &= y_j\mu_{.j} = 0 \qquad (j = 1, \cdots, n), \end{aligned} \tag{5.3}$$

must be satisfied. We shall refer to equations of the form (5.2) as equations of a stationary state. And, if the stationary state is non-trivial, i.e. all x_i and y_j are not zero, then the equations

$$\begin{aligned} \lambda_{i.} &= 0 \qquad (i = 1, \cdots, m), \\ \mu_{.j} &= 0 \qquad (j = 1, \cdots, n), \end{aligned} \tag{5.4}$$

must be satisfied.

By writing equations (5.4) out in full, we see that the stationary state vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ must satisfy the algebraic equations,

$$\mathbf{B}'\hat{\mathbf{x}} = \mathbf{0}_1, \qquad \mathbf{A}\hat{\mathbf{y}} = \mathbf{0}_2, \tag{5.5}$$

where $\mathbf{0}_1$ and $\mathbf{0}_2$ are $n \times 1$ and $m \times 1$ zero vectors respectively. By continuing in this way, we may find the defining equations, the equations of a stationary state, and a set of equivalent algebraic equations of the three remaining stationary state systems. The results are given in Table 1.

For the case of the matrices **A** and **B** nonconstant, we may again write down the defining equations of a class, the equations of a stationary state, and the algebraic equations of a stationary state. We shall, however, restrict our considerations to that class of fitness functions which are such that the algebraic equations of a stationary state admit at least one nontrivial solution.

6. SOLUTIONS OF THE STATIONARY STATE EQUATIONS

In this section we shall state some conditions under which solutions of the stationary state equations exist for the case **A** and **B** are constant matrices. For a detailed treatment of the methods used in this section the reader is referred to a book on matrix algebra such as Perlis [1952]. We shall say a stationary state is unique if, and only if, the equations of a stationary state admit a unique solution. The four classes of systems will be considered in order.

The algebraic equations for a stationary state system of Class I are

$$\mathbf{B}'\hat{\mathbf{x}} = \mathbf{0}_1 \,, \tag{6.1}$$

$$\mathbf{A}\hat{\mathbf{y}} = \mathbf{0}_2 \,. \tag{6.2}$$

Note that (6.1) is a set of n linear equations in m unknowns and (6.2) is a set of m linear equations in n unknowns. Let min (m, n) be the minimum of m and n. If the rank of **B** is r_1 and that of **A** is r_2, then neither r_1 nor r_2 can exceed min (m, n). Moreover, equations (6.1) and (6.2) admit a nontrivial solution if, and only if, the relations $r_1 < m$ and $r_2 < n$ are satisfied.

Let $\mathcal{S}_1$ be the set of all $\hat{\mathbf{x}}$ satisfying equation (6.1) and let $\mathcal{S}_2$ be the set of all $\hat{\mathbf{y}}$ satisfying equation (6.2). The sets $\mathcal{S}_1$ and $\mathcal{S}_2$ are vector spaces. Thus, any multiple of a vector or linear combination of vectors belonging to the set is again a member of the set.

If the rank of **B** is $r_1 < m$, then equation (6.1) may be reduced by elementary row operations to the form

$$\left[\begin{array}{c:c} \mathbf{I}_{r_1} & \mathbf{B}_1 \\ \hdashline \multicolumn{2}{c}{\mathbf{0}_3} \end{array}\right]\hat{\mathbf{x}} = \mathbf{0}_1 \,, \tag{6.3}$$

where $\mathbf{I}_{r_1}$ is an identity matrix of order r_1, $\mathbf{B}_1$ is a $r_1 \times (m - r_1)$ matrix of constants, and $\mathbf{0}_3$ is a $n - r_1 \times m$ zero matrix. The row vectors of

TABLE 1

Classes of Stationary State Systems

I	II	III	IV
	Defining Equations[1]		
$\mathring{H} = \lambda_{..} = 0$ $\mathring{P} = \mu_{..} = 0$	$\mathring{H} = \lambda_{..} = 0$ $\mathring{P} = \mu_{..} \neq 0$	$\mathring{H} = \lambda_{..} \neq 0$ $\mathring{P} = \mu_{..} = 0$	$\mathring{H} = \lambda_{..} \neq 0$ $\mathring{P} = \mu_{..} \neq 0$
	Equations of a Stationary State[2]		
$dx_i/dt = x_i\lambda_{i.} = 0$ $dy_j/dt = y_i\mu_{.j} = 0$	$dx_i/dt = x_i\lambda_{i.} = 0$ $dy_j/dt = y_j(\mu_{.j} - \mu_{..}) = 0$	$dx_i/dt = x_i(\lambda_{i.} - \lambda_{..}) = 0$ $dy_j/dt = y_j\mu_{.j} = 0$	$dx_i/dt = x_i(\lambda_{i..} - \lambda_{..}) = 0$ $dy_j/dt = y_j(\mu_{.j} - \mu_{..}) = 0$
	Equivalent Algebraic Equations of a Stationary State		
[3] $\mathbf{B}'\hat{\mathbf{x}} = \mathbf{0}_1$ $\mathbf{A}\hat{\mathbf{y}} = \mathbf{0}_2$	$\mathbf{B}'\hat{\mathbf{x}} = \mathbf{c}_2$ $\mathbf{A}\hat{\mathbf{y}} = \mathbf{0}_2$	$\mathbf{B}'\hat{\mathbf{x}} = \mathbf{0}_1$ $\mathbf{A}\hat{\mathbf{y}} = \mathbf{c}_1$	$\mathbf{B}'\hat{\mathbf{x}} = \mathbf{c}_2$ $\mathbf{A}\hat{\mathbf{y}} = \mathbf{c}_1$

[1] $\dot{H} = d(\log H)/dt$, $\mathring{P} = d(\log P)/dt$.

[2] $i = 1, \cdots, m$ and $j = 1, \cdots, n$.

[3] $\mathbf{B}'$ represents the transpose of the matrix $\mathbf{B}$.

the $(m - r_1) \times m$ matrix

$$(\mathbf{B}_1' \mid -\mathbf{I}_{m-r_1}) \tag{6.4}$$

are solutions of equation 6.1 and form a basis for the vector space $\mathcal{S}_1$.

Similarly, if the rank of $\mathbf{A}$ is $r_2 < n$, then (6.2) may be reduced to the form

$$\left[\begin{array}{c|c} \mathbf{I}_{r_2} & \mathbf{A}_1 \\ \hline \multicolumn{2}{c}{\mathbf{0}_4} \end{array}\right] \hat{\mathbf{y}} = \mathbf{0}_2 \tag{6.5}$$

where $\mathbf{I}_{r_2}$ is an identity matrix of order r_2 , $\mathbf{A}_1$ is a $r_2 \times (n - r_2)$ matrix of constants, and $\mathbf{0}_4$ is a $(m - r_2) \times n$ zero matrix. The row vectors of the $(n - r_2) \times n$ matrix

$$(\mathbf{A}_1' \mid -\mathbf{I}_{n-r_2}) \tag{6.6}$$

are solutions of equations (6.2) and form a basis for the vector space $\mathcal{S}_2$.

Any pair of vectors $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ where $\hat{\mathbf{x}}$ belongs to $\mathcal{S}_1$ and $\hat{\mathbf{y}}$ belongs to $\mathcal{S}_2$ with coordinates satisfying the conditions, $0 \leq \hat{x}_i \leq 1$, $0 \leq \hat{y}_i \leq 1$, $\sum_i \hat{x}_i = 1$, and $\sum_j \hat{y}_j = 1$, is a solution of equations (6.1) and (6.2) and are, therefore, probability vectors of a stationary state. Clearly, the pair $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is not unique so that there exists no unique stationary state for systems of Class I.

For systems of Class II the algebraic equation of a stationary state are

$$\mathbf{B}'\hat{\mathbf{x}} = \mathbf{c}_2 \ , \tag{6.7}$$

$$\mathbf{A}\hat{\mathbf{y}} = \mathbf{0}_2 \ . \tag{6.8}$$

Equation (6.8) may be solved by the methods discussed in Class I systems but (6.7) needs special consideration. Writing (6.7) componentwise we have

$$\sum_{i=1}^{m} b_{ji}\hat{x}_i = \sum_{i=1}^{m} \hat{x}_i\hat{\mu}_{i.} \qquad (j = 1, \cdots, n), \tag{6.9}$$

where

$$\mu_{i.} = \sum_{j=1}^{n} b_{ij}\hat{y}_j \ .$$

For each probability vector $\hat{\mathbf{y}}$ which is a solution of (6.8), we may find a $\hat{\mu}_{i.}$ which is independent of $\hat{\mathbf{x}}$, and for each $\hat{\mu}_{i.}$ corresponding to a $\hat{\mathbf{y}}$ we have the system of homogeneous equations

$$\sum_{i=1}^{m} (b_{ji} - \hat{\mu}_{i.})x_i = 0 \qquad (j = 1, \cdots, n). \tag{6.10}$$

It is clear that the methods discussed in Class I systems may again be applied to find solutions of equations (6.10). We note that there exists no unique stationary state for Class II systems since there is no unique solution of equations (6.7) and (6.8). It is clear that the methods used to find stationary state solutions for Class II systems may again be applied to Class III systems.

For systems of class IV the algebraic equations of a stationary state are

$$\mathbf{B}'\hat{\mathbf{x}} = \mathbf{c}_2 \,, \tag{6.11}$$

$$\mathbf{A}\hat{\mathbf{y}} = \mathbf{c}_1 \,. \tag{6.12}$$

Equations (6.11) and (6.12) cannot be solved by the methods used heretofore since they are nonlinear in the components of the unknown vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. If at least one of the vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ is constant, then we may use the methods of solving linear equations to find solutions of (6.11) and (6.12).

For example, if the vector $\mathbf{c}_2$ is constant, then equation (6.11) admits a solution if, and only if, the rank of the matrix $\mathbf{B}'$, satisfies the relation $r_1 \leq m$ and the rank of the augmented matrix $(\mathbf{B}', \mathbf{c}_2)$ is r_1. If $r_1 < m$, then (6.11) admits infinitely many solutions. If $r_1 = m$, then (6.11) admits a unique solution which may be found by Cramer's rule. In all cases we may place conditions on the elements of the matrix $\mathbf{B}'$ so that at least one probability vector is a solution of (6.11).

For each probability vector $\hat{\mathbf{x}}$ which is a solution of (6.11) it may be shown by the methods used in the discussion of Class II systems that (6.12) becomes a homogeneous equation in $\hat{\mathbf{y}}$. In general, homogeneous equations do not admit unique nontrivial solutions so that a unique stationary state cannot exist in this case. The cases in which $\mathbf{c}_1$ is constant and both $\mathbf{c}_1$ and $\mathbf{c}_2$ are constant may be treated similarly.

It is of interest to note that if (1) the vectors $\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_k$ are solutions of (6.11), (2) the vectors $\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_k$, are solutions of (6.12), and and (3) a_i $(i = 1, \cdots, k)$ are positive numbers which satisfy the condition, $a_1 + \cdots + a_k = 1$, then $a_1\hat{\mathbf{x}}_1 + \cdots + a_k\hat{\mathbf{x}}_k$ and $a_1\hat{\mathbf{y}}_1 + \cdots + a_k\hat{\mathbf{y}}_k$ are also solutions of (6.11) and (6.12) respectively.

The case in which $m = n$, $\mathbf{A}$ and $\mathbf{B}$ are nonsingular, and the host and pathogen are associated at random is covered in the following interesting theorem:

Theorem 6.1: If (1) $m = n$, (2) the matrices $\mathbf{A}$ and $\mathbf{B}$ are nonsingular, and (3) all $\theta_{ij} = 1$, then $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\lambda}_{..}$, and $\hat{\mu}_{.}$ are unique so that a unique stationary state may exist in systems of Class IV.

Proof: Let $\hat{\lambda}_{..}$ and $\hat{\mu}_{..}$ be the values of $\lambda_{..}$ and $\mu_{..}$ corresponding to the stationary state vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. If $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are stationary-state vectors, then their components, by Cramer's rule, must satisfy the equations.

$$\begin{aligned} \hat{x}_i &= \hat{\mu}_{..} \mid \mathbf{B}_i \mid / \mid \mathbf{B} \mid \qquad (i = 1, \cdots, m), \\ \hat{y}_j &= \hat{\lambda}_{..} \mid \mathbf{A}_j \mid / \mid \mathbf{A} \mid \qquad (j = 1, \cdots, m), \end{aligned} \tag{6.13}$$

where $\mathbf{B}_i$ is matrix obtained from $\mathbf{B}$ by replacing the i-th row by a row of ones, $\mathbf{A}_j$ is a matrix obtained from $\mathbf{A}$ by replacing the j-th column by a column of ones, and $\mid \mathbf{B}_i \mid$, $\mid \mathbf{A}_j \mid$, $\mid \mathbf{B} \mid$, and $\mid \mathbf{A} \mid$ are the determinants of the matrices in question. Summing equations (6.13) over i and j and using the condition the components of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ sum to one yields

$$\begin{aligned} \hat{\lambda}_{..} &= \mid \mathbf{A} \mid / \sum_{j=1}^{m} \mid \mathbf{A}_j \mid , \\ \hat{\mu}_{..} &= \mid \mathbf{B} \mid / \sum_{i=1}^{m} \mid \mathbf{B}_i \mid . \end{aligned} \tag{6.14}$$

By substituting $\hat{\lambda}_{..}$ and $\hat{\mu}_{..}$ in (6.13) we find

$$\begin{aligned} \hat{x}_i &= \mid \mathbf{B}_i \mid / \sum_{i=1}^{m} \mid \mathbf{B}_i \mid , \\ \hat{y}_i &= \mid \mathbf{A}_j \mid / \sum_{j=1}^{m} \mid \mathbf{A}_j \mid . \end{aligned} \tag{6.15}$$

To prove uniqueness, let $\lambda_{..}^*$ and $\mu_{..}^*$ be the values of $\lambda_{..}$ and $\mu_{..}$ corresponding to the stationary state vectors $\mathbf{x}^*$ and $\mathbf{y}^*$. Proceeding as before, we have by Cramer's rule

$$\begin{aligned} x_i^* &= \mu_{..}^* \mid \mathbf{B}_i \mid / \mid \mathbf{B} \mid , \\ y_i^* &= \lambda_{..}^* \mid \mathbf{A}_j \mid / \mid \mathbf{A} \mid . \end{aligned} \tag{6.16}$$

By summing over i and j and using the condition the components of $\mathbf{x}^*$ and $\mathbf{y}^*$ sum to one we find $\hat{\lambda}_{..} = \lambda_{..}^*$ and $\hat{\mu}_{..} = \mu_{..}^*$, and by substituting $\hat{\lambda}_{..}$ and $\hat{\mu}_{..}$ in (6.16) we reach the conclusion that $\hat{\mathbf{x}} = \mathbf{x}^*$ and $\hat{\mathbf{y}} = \mathbf{y}^*$ which proves the theorem.

It will be noted that in order that $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ be probability vectors we must require that the elements of the matrices $\mathbf{A}$ and $\mathbf{B}$ be such that the components of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ be non-negative. It will also be noted that by dropping the assumptions of equal numbers of varieties and races and random association of the host and pathogen, we are led to the possible existence of non-unique stationary states.

7. THE STABILITY OF A STATIONARY STATE

We next turn to the question of stability of a stationary state corresponding to a pair of stationary state vectors $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$. For brevity, we will simply refer to the stability of a pair $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$.

Let $\mathbf{z}' = (\mathbf{x}', \mathbf{y}')$ be a $1 \times (m + n)$ vector and rename the components $z_k(k = 1, \cdots, m + n)$. Let $\mathbf{dz}/\mathbf{dt}$ be a column vector with components dz_k/dt, and let $\mathbf{u}(\mathbf{z})$ be a column vector with components $u_k(\mathbf{z}) = x_k(\lambda_{k.} - \lambda_{..})$ for $k = 1, \cdots, m$ and $u_k(\mathbf{z}) = y_k(\mu_{.k} - \mu_{..})$ for $k = m + 1, \cdots, m + n$. with these definitions, differential equations (2.8) may be written in the compact form

$$\mathbf{dz}/\mathbf{dt} = \mathbf{u}(\mathbf{z}). \tag{7.1}$$

In this paper the definition of stability given by Bellman [1953, p. 76] will be used. Using this definition, the stability of a stationary vector $\hat{\mathbf{z}}$ may be decided by using the following procedure. Let $\mathbf{h}' = \mathbf{z}' - \hat{\mathbf{z}}' = (h_1, \cdots, h_{m+n})$. Now by the multivariate version of Taylor's theorem, we may express $\mathbf{u}(\mathbf{z})$ in the form

$$\mathbf{u}(\mathbf{z}) = \mathbf{u}(\hat{\mathbf{z}}) + \mathbf{Qh} + \mathbf{v}(\mathbf{h}), \tag{7.2}$$

where $\mathbf{Q}$ is the $(m + n) \times (m + n)$ Jacobian matrix of $\mathbf{u}(\mathbf{z})$ evaluated at $\hat{\mathbf{z}}$ and $\mathbf{v}(\mathbf{h})$ stands for a vector of nonlinear terms. Clearly, $\mathbf{dh}/\mathbf{dt} = \mathbf{dz}/\mathbf{dt}$ since $\hat{\mathbf{z}}$ is a constant vector. Moreover, if $\hat{\mathbf{z}}$ is a stationary state vector, then $\mathbf{u}(\hat{\mathbf{z}})$ equals the zero vector. The question of the stability of vector $\hat{\mathbf{z}}$ is thus reduced to question of the stability of the trivial solution $\mathbf{h}' = (0, \cdots, 0)$ of the differential equation

$$\mathbf{dh}/\mathbf{dt} = \mathbf{Qh} + \mathbf{v}(\mathbf{h}). \tag{7.3}$$

We shall obtain conditions for stability under the assumption differential equation (7.3) satisfies the hypotheses of Theorem 1, page 79 of Bellman [1953]. Before this theorem may be used, however, the following observations are essential. From the definition of the $\mathbf{h}$, it is easy to see its components satisfy the conditions

$$\sum_{k=1}^{m} h_k = 0 \quad \text{and} \quad \sum_{k=m+1}^{m+n} h_k = 0. \tag{7.4}$$

Therefore, any meaningful solution of differential equation (7.3) must also satisfy conditions (7.4). It may be easily shown, although we shall not do so here, that, if the initial conditions satisfy conditions (7.4), then all solutions of (7.3) satisfy conditions (7.4). This result permits us to work directly with the Jacobian matrix $\mathbf{Q}$.

One of the hypothesis of the stability theorem (Theorem 1, p. 79, Bellman) is that all solutions of the differential equation

$$\mathbf{dh/dt} = \mathbf{Qh}, \tag{7.5}$$

approach zero as $t \to \infty$. All solutions of (7.5) approach zero as $t \to \infty$ if, and only if, the real parts of the characteristic roots of $\mathbf{Q}$ are negative. (Theorem 7, p. 25, Bellman.) The stability of a vector $\hat{\mathbf{z}}$ is, therefore, determined by the properties of the matrix $\mathbf{Q}$.

Let us next examine the structure of the matrix $\mathbf{Q}$. The matrix $\mathbf{Q}$ is such that it has the following form

$$\mathbf{Q} = \left[\begin{array}{c:c} \mathbf{Q}_1 & \mathbf{Q}_2 \\ \hdashline \mathbf{Q}_3 & \mathbf{Q}_4 \end{array}\right], \tag{7.6}$$

and the submatrices $\mathbf{Q}_1$, $\mathbf{Q}_2$, $\mathbf{Q}_3$, and $\mathbf{Q}_4$ have the following forms: $\mathbf{Q}_1$ is a $m \times m$ matrix with diagonal elements

$$\hat{q}_{1ii} = \hat{x}_i\left(\widehat{\frac{\partial\lambda_{i.}}{\partial x_i}} - \widehat{\frac{\partial\lambda_{..}}{\partial x_i}}\right) \qquad (i = 1, \cdots, m), \tag{7.7}$$

and nondiagonal elements

$$\hat{q}_{1ii'} = \hat{x}_i\left(\widehat{\frac{\partial\lambda_{i.}}{\partial x_{i'}}} - \widehat{\frac{\partial\lambda_{..}}{\partial x_{i'}}}\right) \qquad (i \neq i'). \tag{7.8}$$

$\mathbf{Q}_2$ is a $m \times n$ matrix whose ji-th element is

$$\hat{q}_{2ij} = \hat{x}_i\left(\widehat{\frac{\partial\lambda_{i.}}{\partial y_j}} - \widehat{\frac{\partial\lambda_{..}}{\partial y_j}}\right) \qquad (j = 1, \cdots, n). \tag{7.9}$$

$\mathbf{Q}_3$ is a $n \times m$ matrix whose ji-th element is

$$\hat{q}_{3ji} = \hat{y}_j\left(\widehat{\frac{\partial\mu_{.j}}{\partial x_i}} - \widehat{\frac{\partial\mu_{..}}{\partial x_i}}\right). \tag{7.10}$$

$\mathbf{Q}_4$ is a $n \times n$ matrix with diagonal elements

$$\hat{q}_{4jj} = \hat{y}_j\left(\widehat{\frac{\partial\mu_{.j}}{\partial y_j}} - \widehat{\frac{\partial\mu_{..}}{\partial y_j}}\right), \tag{7.11}$$

and nondiagonal elements

$$\hat{q}_{4jj'} = \hat{y}_j\left(\widehat{\frac{\partial\mu_{.j}}{\partial y_{j'}}} - \widehat{\frac{\partial\mu_{..}}{\partial y_{j'}}}\right). \tag{7.12}$$

Recall that in the above expressions the carats stand for the evaluation of the elements of the Jacobian matrix of $\mathbf{u}(\mathbf{z})$ at a stationary state vector $\hat{\mathbf{z}}' = (\hat{\mathbf{x}}', \hat{\mathbf{y}}')$.

The following equations are essential for the determination of the properties of the matrix $\mathbf{Q}$.

$$\frac{\partial \lambda_{i.}}{\partial x_i} = \sum_{j=1}^{n} y_j \theta_{ij} \frac{\partial \lambda_{ij}}{\partial x_i} ; \qquad \frac{\partial \lambda_{..}}{\partial x_i} = \lambda_{i.} + x_i \frac{\partial \lambda_{i.}}{\partial x_i} , \tag{7.13}$$

$$\frac{\partial \lambda_{i.}}{\partial y_j} = \theta_{ij} \lambda_{ij} + y_j \theta_{ij} \frac{\partial \lambda_{ij}}{\partial y_j} ; \qquad \frac{\partial \lambda_{..}}{\partial y_j} = \sum_{i=1}^{m} x_i \frac{\partial \lambda_{i.}}{\partial y_j} . \tag{7.14}$$

Expressions for $\partial\mu_{.j}/\partial x_i$, $\partial\mu_{..}/\partial x_i$, $\partial\mu_{.j}/\partial y_j$, and $\partial\mu_{..}/\partial y_j$ are similar to (7.13) and (7.14) and may be written down from symmetry considerations.

It is of interest to examine stability in the following five cases. The truth of the statements to follow can easily be deduced from equations (7.7) through (7.14).

1. If a system belongs to Class I and all λ_{ij} and μ_{ij} are constant, then the Jacobian matrix **Q** has the form

$$\left[\begin{array}{c:c} \mathbf{0} & \mathbf{Q}_2 \\ \hdashline \mathbf{Q}_3 & \mathbf{0} \end{array}\right], \tag{7.15}$$

where **0** stands for a square zero matrix. Now from the theory of characteristic roots it is known that the sum of the elements on the principal diagonal equals the sum of the characteristics roots. In this case the sum of the elements on the principal diagonal is zero. hence, the sum of the characteristics roots is zero. It follows that all real parts of the characteristics roots cannot have the same sign and, therefore, cannot be negative. Thus, no stable stationary states can exist in this case. Note the role constant population numbers in the host and pathogen populations and constant fitness functions play in the instability of this case.

2. If a system belongs to Class II and all λ_{ij} and μ_{ij} are constant, then the matrix **Q** has the form

$$\left[\begin{array}{c:c} \mathbf{0} & \mathbf{Q}_2 \\ \hdashline \mathbf{Q}_3 & \mathbf{Q}_4 \end{array}\right]. \tag{7.16}$$

3. If a system belongs to Class III and all λ_{ij} and μ_{ij} are constant, then the matrix **Q** has the form

$$\left[\begin{array}{c:c} \mathbf{Q}_1 & \mathbf{Q}_2 \\ \hdashline \mathbf{Q}_3 & \mathbf{0} \end{array}\right]. \tag{7.17}$$

4. If the system belongs to Class IV and all λ_{ij} and μ_{ij} are constant, then the matrix **Q** will have no submatrices which are necessarily a zero matrix.

5. If all λ_{ij} and μ_{ij} are nonconstant, then the matrix **Q** will not have a submatrix which is necessarily a zero matrix. This statement is true of systems of Class I, II, III and IV.

The following theorem supplies a sufficient condition for stability in cases 2, 3, 4 and 5.

Theorem 7.1: Construct a matrix **Q*** from **Q** as follows. Let $q^*_{kk} = q_{kk}$ $(k = 1, \cdots, m + n)$ and set $q^*_{kk'} = \frac{1}{2}(q_{kk'} + q_{k'k})$ for $k \neq k'$. If **Q*** is negative definite, then the real parts of the characteristics roots of **Q** are negative.

Proof: Let λ be a complex characteristic root of Q and let $\boldsymbol{\beta}_1 + i\boldsymbol{\beta}_2$ $(i^2 = -1)$ be its associated complex characteristic vector of dimensions $(m + n) \times 1$. From the definition of characteristic roots and vector we have

$$\mathbf{Q}(\boldsymbol{\beta}_i + i\boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)\lambda. \tag{7.18}$$

Now multiply equation (7.18) by $(\boldsymbol{\beta}_i - \mathbf{i}\boldsymbol{\beta}_2)'$, the transpose of the complex conjugate of $\boldsymbol{\beta}_1 + \mathbf{i}\boldsymbol{\beta}_2$. The result is

$$\lambda = (\boldsymbol{\beta}_1'\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2'\boldsymbol{\beta}_2)^{-1}(\boldsymbol{\beta}_1'\mathbf{Q}\boldsymbol{\beta}_i + \boldsymbol{\beta}_2\mathbf{Q}\boldsymbol{\beta}_2 + i(\boldsymbol{\beta}_1'\mathbf{Q}\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2'\mathbf{Q}\boldsymbol{\beta}_1)). \tag{7.19}$$

Thus since $\boldsymbol{\beta}_1'\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2'\boldsymbol{\beta}_2$ is positive the sign of the real part of a characteristic root depends on the sign of $\boldsymbol{\beta}_1'\mathbf{Q}\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2'\mathbf{Q}\boldsymbol{\beta}_2$. But $\boldsymbol{\beta}_1'\mathbf{Q}\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2'\mathbf{Q}\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1'\mathbf{Q}^*\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2'\mathbf{Q}^*\boldsymbol{\beta}_2$. It follows that if the matrix **Q*** is negative definite, then the real parts of all characteristic roots will be negative, which completes the proof of the theorem.

In cases 2, 3, 4 and 5 it is possible to construct a matrix **Q*** from **Q** so that **Q*** is negative definite. Therefore, stable stationary states may exist in cases 2, 3, 4 and 5.

8. INTERPRETATIONS AND PRACTICAL CONSIDERATIONS.

A little consideration will lead to the conclusion that the class of stationary state system to which a given host-pathogen system may belong depends on the length of the time interval and the size of the geographical area under consideration. For example, suppose our host-pathogen system is a given cereal crop and some species of rust. If we consider this system with respect to the time interval, $0 \leq t \leq t_1$, representing a single growing season and a single field of the crop, then any stationary state system would probably belong to Class II; since in a given growing season the number of members of the host population is essentially constant but the number of members of the pathogen population would probably be increasing. On the other-hand, if the time interval is taken to be some period of years and a larger geographical area were considered, then our host-pathogen system would probably belong to Class IV; since during this period of years the number of members of both the host and pathogen populations would probably be changing. It seems likely that the majority of host-pathogen systems encountered in practice belong to systems of Class IV.

A question that arises is what procedure should one follow if he wishes to construct a host-pathogen system so that (1) it is in a stable stationary state, and (2) the number of members in the pathogen population at the end of any growing season is less some critical number c. In general, the parameters of the system can never be known exactly. At the present time, therefore, the most fruitful approach seems to be an experimental one. That is, simply construct a set of systems consisting of a mixture of varieties and races and observe their behavior over a period of time. Any system satisfying the above conditions in this time interval would, apparently, be satisfactory from the practical point of view.

Some suggestions for the construction of a host-pathogen systems meeting conditions (1) and (2) of the above paragraph for the case of equal numbers of varieties and races, constant fitness functions, and random association of the host and pathogen are given in the previous paper, Mode [1960].

9. SUMMARY

The results of this paper represent a generalization of results given in a previous paper. In this paper a characterization of a host-pathogen system containing an arbitrary number of host varieties and races of the pathogen was given under the assumptions of nonrandom association of the host and pathogen and nonconstant fitness functions.

Four classes of stationary state systems were defined on the basis of constant or nonconstant population numbers in the host and pathogen populations. Some methods of finding probability vectors of a stationary state were given for the four classes of systems.

The four classes of stationary state systems were also checked for stability. It was found that if the fitness functions are constant, then a stable stationary state may exist in systems of Class II, III, or IV but not in systems of Class I. If the fitness functions are not constant, then it is possible for a stable stationary state to exist in any of the four classes of systems. The possible existence of stable stationary states is of considerable practical interest.

REFERENCES

Bellman, R. [1953]. *Stability Theory of Differential Equations.* New York. McGraw-Hill.

Flor, H. H. [1956]. The complementary genic systems in flax and flax rust. New York. Academic Press Inc. *Advances In Genetics*, 29–54.

Kimura, M. [1958]. On the change of population fitness by natural selection. *Heredity* 12, 145–67.

Mode, C. J. [1960]. A model of a host-pathogen system with particular reference to the rusts of cereals. New York. Pergamon Press. *Biometrical Genetics*, 84–96.

Perlis, S. [1952]. *Theory of Matrices.* Reading, Mass., Addison-Wesley.

Suneson, C. A. [1960]. Genetic diversity—A protection against plant diseases and, insects. *Agronomy Journal* 52, 318–21.

LATIN SQUARES TO BALANCE IMMEDIATE RESIDUAL, AND OTHER ORDER, EFFECTS.

Paul R. Sheehe and Irwin D. J. Bross
Roswell Park Memorial Institute
Buffalo, New York, U.S.A.

1. INTRODUCTION

While the Latin Square designs to be presented have a broad field of application, the focus of our discussion will be on their use in clinical trials. These extra-balanced designs were, in fact, developed specifically to meet an experimental problem that arose in setting up an analgesic trial. This specific problem will be considered briefly in order to illustrate the scientific (as distinguished from purely mathematical) rationale for doing balanced experiments.

The original study plan for the analgesic trial called for five agents to be used: a placebo, a standard drug, a new agent, a combination of the new and standard agents, and a course of hypnosis. Each patient was to be his own control, but a decision had to be made about the duration of each treatment. If the duration of each treatment were as long as a week, drop-outs, with resulting incomplete sequences of treatments, would become a serious practical problem. In view of this, the principal investigator considered a shorter trial period for each agent, such as three days, to be preferable. He was confident that this period of observation would be long enough to elicit reliable responses and, so far as straight chemical carry-over, or residual, effects were concerned, analgesic effects would last only a few hours. But then he recalled a kind of psychological carry-over effect that he had noticed in previous studies. When an effective agent was given after a placebo or ineffective agent, it seemed that the effective agent often failed. A plausible hypothesis was that the patient had lost confidence in the analgesics and it would take some time for his confidence to be restored. This applied especially to double-blind studies in which the patient often was under the impression that he received the same agent all of the time. At this point in the planning, the investigator therefore wondered if there was some way to insure that each agent tested would be immediately preceded by the placebo an equal number of times. Also, since it was possible that several treatments would be, like placebos, relatively ineffective, he wondered if there was a way to achieve more complete balance for immediate residual effects by insuring that every treatment would be immediately preceded equally often by every other treatment.

The design which achieved this objective was worked out by trial and error, but then the study plan was modified to include two new treatments. This change emphasized the desirability of having a general procedure for constructing predecessor-balanced designs. In the literature, E. J. Williams [1] originally set down, in 1949, the conditions sufficient to produce Latin Squares balanced for immediate predecessors. Sufficient conditions for more remote predecessors, particularly the next-to-last predecessors, were also specified. Williams also presented analysis of variance procedures to accompany the designs. Raymond, et al, [2] used this type of design and analysis in 1957 for a study of tranquillizing drugs in psychoneurosis. In 1952, H. D. Patterson [3] considered the more general problem of predecessor-balanced designs, not only for square arrangements such as Williams', but also for certain incomplete block arrangements.

In 1958, J. V. Bradley [4] presented an easily remembered construction which meets the conditions set down by Williams when the number of treatments is even. He also presented additional balancing procedures which might be useful in special cases. To round out the picture, we shall present here an easily remembered procedure which can be used whether the number of treatments is odd or even. In addition to the Latin Square and immediate predecessor-balance features (properties 1 and 3, resp., to follow), another balance property (property 2, to follow) will be noted. We shall discuss the application of this, and other extra-balanced designs, in the context of the clinical trial. In the appendix, we shall present a detailed proof of the method. A second section of the appendix will be devoted to describing a simple variation in construction which produces Graeco-Latin Squares when the number of treatments is odd.

2. PROCEDURE

Before proceeding further, a more formal definition of 'balance' is appropriate. In addition, some predecessor balance properties, other than immediate, will be defined. A design is called balanced with respect to the set of immediate predecessors if every treatment is immediately preceded equally often by every other treatment. Similarly, a design may be called balanced with respect to the set of predecessors of any specified degree of remoteness (e.g., the second degree, which involves the next-to-last predecessors), if every treatment is preceded equally often by every other treatment at that degree of remoteness. Balance with respect to the set of all predecessors (without regard to degree of remoteness) is achieved if every treatment is preceded, immediately or more remotely, equally often by every other treatment. The term, complete balance, is reserved for the case when the set of predecessors (of a specified or unspecified degree of remoteness) is the same for every treatment. Thus, balanced designs are not completely balanced only because treatments do not precede themselves. The design presented here is (1) completely balanced with respect to the number of treatments preceding every treatment, (2) balanced with respect to the set of all predecessors, and (3) balanced with respect to the set of immediate predecessors.

The procedure for construction is as follows:

(a) Number the treatments, $i = 1, \cdots, n$.

(b) Start with a cyclic $n \times n$ Latin Square, i.e. one in which the sequence of treatments in the ith row is $i, i + 1, \cdots, n, 1, 2, \cdots, i - 1$.

(c) Interlace each row of the cyclic Latin Square with its own reverse order sequence, i.e. with its mirror image. For example, if $n = 5$, the first row of the cyclic Latin Square reads 1, 2, 3, 4, 5. Its mirror image is 5, 4, 3, 2, 1, and when this is interlaced with the first row of the original square, the interlaced sequence reads 1, 5, 2, 4, 3, 3, 4, 2, 5, 1.

(d) Slice the resulting $n \times 2n$ figure down the middle, thus forming two $n \times n$ Latin Squares. The columns of each square refer to the order of presentation, from left to right, and the rows refer to individuals. Treatments appear in the body of each square.

It will be found that, when n is even, each of the constructed squares has the three desired properties. In this case, either of the two squares may be used. When n is odd, each of the constructed squares has the first property, but not the last two. However, when the two squares are considered as a whole, the last two properties are indeed present. Consequently, in this case, both of the constructed squares must be used. Constructed squares for $n = 4$ and $n = 5$ are presented in Tables 1 and 2. The reader may verify, by inspection, that the stated properties are present in either case. It will also be noted that, for $n = 4$ or for any even number, in general, the left square is identical with that originally presented by Bradley. Proof that all these properties hold in general will be offered in the appendix.

3. DISCUSSION

Lest the reader be left with the impression that, because the Latin Square principle has been used to achieve a desired order balance, a Latin Square analysis is advised, we specifically disclaim this as our

TABLE 1

Constructed Latin Squares for $n = 4$

Left Square					Right Square				
	Order of Presentation					Order of Presentation			
Individual	1	2	3	4	Individual	1	2	3	4
A	1	4	2	3	E	3	2	4	1
B	2	1	3	4	F	4	3	1	2
C	3	2	4	1	G	1	4	2	3
D	4	3	1	2	H	2	1	3	4

intention. In the context of the clinical trial where the problem of predecessor balance arose, the primary function of the balancing was not to reduce experimental error, but to provide a safeguard against a fairly specific danger. This safeguard is present in the designs whether we regard them as Latin Squares, or whether we regard individuals as blocks. Also, the main reason for using the patient as his own control in an analgesic trial is that the response variable is in subjective scale, dependent on the value judgment of the individual patient. Thus, if a fairly large number of individuals were available for the trial, a plausible method of analysis might be to analyze intra-patient comparisons. In fact, the nature of the response variable is not very different from that encountered in paired comparison trials. Furthermore, the designs here are such that, if each treatment is paired with its immediate predecessor, all possible pairs of different treatments will be formed equally often. This suggests that an adaptation of the paired

TABLE 2

Constructed Latin Squares for $n = 5$

Left Square						Right Square					
	Order of Presentation						Order of Presentation				
Individual	1	2	3	4	5	Individual	1	2	3	4	5
A	1	5	2	4	3	F	3	4	2	5	1
B	2	1	3	5	4	G	4	5	3	1	2
C	3	2	4	1	5	H	5	1	4	2	3
D	4	3	5	2	1	I	1	2	5	3	4
E	5	4	1	3	2	J	2	3	1	4	5

comparison analysis presented by Scheffé [5] might be satisfactory for practical purposes. Cochran and Cox [6] give an illustrative Latin Square analysis (in the manner of Williams), complicated because of balance for immediate residual effects, but they too caution against the indiscriminate use of the method when the underlying assumptions are suspect.

In some clinical situations, neither a Latin Square nor a one-way analysis of all the data would be advisable. For example, in a post-operative trial of analgesics, Meier et al. [7] report that observations made in the first few post-operative hours, when patient pain status was most acute, were much more discriminatory for the efficacy of the agents than were observations made thereafter. If, as was actually done in the cited case, the later observations were dropped from the analysis, the remaining data would still be at least partially balanced. Thus, regardless of what analysis might be most appropriate, the predecessor balance feature serves as a kind of 'ace-in-the hole'. If, when the study is subsequently reported, the critic should raise the question of psychological or other carry-over effects, the investigator could reply by pointing to the design precaution that had been taken to balance out such effects.

It is most important, however, to remember that predecessor balancing cannot provide perfect protection against carry-over effects. Such effects are not likely to be consistent, that is, they may vary considerably from one patient to another. To control the carry-over effects, the balance would have to be over those patients who exhibited consistent carry-over effects, but unfortunately these sub-sets of patients cannot be distinguished in advance. Nevertheless the control should be closer where there is balance over a single experiment than with the usual Latin Square where there would be balance over a hypothetical large series of experiments. This more modest justification is sufficient for practical purposes. We should not expect any design feature to completely solve a deep-rooted experimental problem—all that can reasonably be asked is that the device improve the degree of control. This point deserves attention because many useful design features have been "over-sold" so that the investigator gets the impression that he is completely covered by his statistical insurance policy.

What are some other limitations and drawbacks of these designs? One limitation is that the designs are not balanced with respect to second-degree (and higher-degree) predecessors. A more practical limitation (for designs with an odd number of treatments) is that the number of patients must be a multiple of $2n$ in order that balance be achieved. For example, with 7 treatments there might be good practical reasons for terminating the experiment after 35 patients or so have come in, but 42 patients would be required to balance the design. This limitation is less important, however, when the emphasis is on precautions rather than analysis.

Another limitation is that complete balance (as defined) with respect to immediate predecessors is not achieved. But by slightly modifying the designs given here, complete balance can be achieved by simply inserting a "zeroth order" column, identical with the first column, in each square. If the data from the zeroth order column are discarded in the analysis, it will be found that each treatment is immediately preceded equally often by every treatment, including itself. Thus the set of immediate predecessors is the same for every treatment. Bradley has already pointed this out in his discussion of an even number of treatments. It is equally true for an odd number of treatments. But the advantages of this further balance would have to be weighed against the possible disadvantages. For example, in the post-operative pain situation reported by Meier et al. and cited above, it would be especially undesirable to discard initial post-operative observations.

Another simple extra-balancing feature, mentioned by Bradley for an even number of treatments but equally applicable for an odd number, provides column balance analogous to the row order balance. The procedure is to permute the rows of each Latin Square in such a way that the first column reads down in the same sequence as the first row reads across. Note, however, that the addition of this feature sharply limits the number of possible Latin Squares. From a theoretical standpoint (e.g. the randomization set justification of the Latin Square analysis) the restriction raises some difficulties. Incidentally this same objection applies, though usually with less force, to any other extra-balanced design. Sir R. A. Fisher has vigorously maintained the position that the analysis of data must take into account the restrictions of the design and this view has been accepted by a majority of statisticians. This point was debated by Student (W. G. Gosset) and Fisher in the 1920's. The discussion of the Knut Vic (Knight's Move) Latin Square is directly relevant here since this square contains an extra-balance feature. In practice this would mean either a least squares analysis or the segregation of individual degrees of freedom associated with the restriction. If the Fisherian view is accepted, then there is a double liability to extra balance. Not only the computational difficulty would be increased but also results would be strongly dependent upon the assumptions of the model (loss of robustness). So we would not advise experimentors to use extra-balanced designs simply as a gimmick (or an "aesthetic" grounds). There should be some sound practical reasons for the additional restrictions.

APPENDIX

Proof

The basis for proving that the three properties listed in Section 2 hold in general is the fact that the left square and the right square are mirror images of each other, i.e. treatments appear symmetrically about the vertical line at which the slice was made in step (d). There is this symmetry by virtue of the method of construction: the first treatment in any row of the left square appears as the last treatment in the corresponding row of the right square, the second in the left appears as the second-last in the right and so on to the last in the left which appears as the first in the right square.

To prove that each of the constructed squares has the first property (complete balance with respect to number of predecessors), whether n is even or odd, it is sufficient to show that each square is, as claimed, a Latin Square. For every Latin Square has this property. By construction, each treatment appears exactly twice in every row of the $n \times 2n$ figure constructed in step (c). By symmetry, every treatment appearing in the left square also appears in the right square, hence each treatment must appear just once in any row of the left square and once in the corresponding row of the right square. Furthermore, the columns of the original cyclic Latin Square remain intact throughout the construction of each square. Thus each constructed square is arrived at, in effect, by permutation of the columns of a cyclic Latin Square. Since any permutation of columns of a Latin Square is also a Latin Square, both left and right figure are Latin Squares. This demonstrates that the first property holds.

It can now be proved that the second property (balance with respect to the set of all predecessors) holds for the two squares taken as a whole. As shown, the two squares are Latin Squares and are mirror images of each other. Consequently, if k different treatments precede (and therefore the remaining $(n - k - 1)$ treatments succeed) a given treatment in a row of the left square then the remaining $(n - k - 1)$ treatments precede (and the k succeed) the given treatment in the mirror image row of the right square. That is to say, in any given row of the two squares taken together, every treatment is preceded (and succeeded) just once by each of the $k + (n - k - 1) = (n - 1)$ other treatments. Taking all n rows into account, every treatment is preceded (and succeeded) n times by each of the $(n - 1)$ other treatments. This constitutes proof that the second property holds when the two constructed squares are taken as a whole. Proof that, when n is even, *each* constructed square has this property will be put off until the third property has been dealt with.

reasons for terminating the experiment after 35 patients or so have come in, but 42 patients would be required to balance the design. This limitation is less important, however, when the emphasis is on precautions rather than analysis.

Another limitation is that complete balance (as defined) with respect to immediate predecessors is not achieved. But by slightly modifying the designs given here, complete balance can be achieved by simply inserting a "zeroth order" column, identical with the first column, in each square. If the data from the zeroth order column are discarded in the analysis, it will be found that each treatment is immediately preceded equally often by every treatment, including itself. Thus the set of immediate predecessors is the same for every treatment. Bradley has already pointed this out in his discussion of an even number of treatments. It is equally true for an odd number of treatments. But the advantages of this further balance would have to be weighed against the possible disadvantages. For example, in the post-operative pain situation reported by Meier et al. and cited above, it would be especially undesirable to discard initial post-operative observations.

Another simple extra-balancing feature, mentioned by Bradley for an even number of treatments but equally applicable for an odd number, provides column balance analogous to the row order balance. The procedure is to permute the rows of each Latin Square in such a way that the first column reads down in the same sequence as the first row reads across. Note, however, that the addition of this feature sharply limits the number of possible Latin Squares. From a theoretical standpoint (e.g. the randomization set justification of the Latin Square analysis) the restriction raises some difficulties. Incidentally this same objection applies, though usually with less force, to any other extra-balanced design. Sir R. A. Fisher has vigorously maintained the position that the analysis of data must take into account the restriction of the design and this view has been accepted by a majority of statisticians. This point was debated by Student (W. G. Gosset) and Fishe in the 1920's. The discussion of the Knut Vic (Knight's Move) Latin Square is directly relevant here since this square contains an extra balance feature. In practice this would mean either a least square analysis or the segregation of individual degrees of freedom associate with the restriction. If the Fisherian view is accepted, then there is double liability to extra balance. Not only the computational difficult would be increased but also results would be strongly dependent upo the assumptions of the model (loss of robustness). So we would nc advise experimentors to use extra-balanced designs simply as a gimmic (or an "aesthetic" grounds). There should be some sound practica reasons for the additional restrictions.

APPENDIX

Proof

The basis for proving that the three properties listed in Section 2 hold in general is the fact that the left square and the right square are mirror images of each other, i.e. treatments appear symmetrically about the vertical line at which the slice was made in step (d). There is this symmetry by virtue of the method of construction: the first treatment in any row of the left square appears as the last treatment in the corresponding row of the right square, the second in the left appears as the second-last in the right and so on to the last in the left which appears as the first in the right square.

To prove that each of the constructed squares has the first property (complete balance with respect to number of predecessors), whether n is even or odd, it is sufficient to show that each square is, as claimed, a Latin Square. For every Latin Square has this property. By construction, each treatment appears exactly twice in every row of the $n \times 2n$ figure constructed in step (c). By symmetry, every treatment appearing in the left square also appears in the right square, hence each treatment must appear just once in any row of the left square and once in the corresponding row of the right square. Furthermore, the columns of the original cyclic Latin Square remain intact throughout the construction of each square. Thus each constructed square is arrived at, in effect, by permutation of the columns of a cyclic Latin Square. Since any permutation of columns of a Latin Square is also a Latin Square, both left and right figure are Latin Squares. This demonstrates that the first property holds.

It can now be proved that the second property (balance with respect to the set of all predecessors) holds for the two squares taken as a whole. As shown, the two squares are Latin Squares and are mirror images of each other. Consequently, if k different treatments precede (and therefore the remaining $(n - k - 1)$ treatments succeed) a given treatment in a row of the left square then the remaining $(n - k - 1)$ treatments precede (and the k succeed) the given treatment in the mirror image row of the right square. That is to say, in any given row of the two squares taken together, every treatment is preceded (and succeeded) just once by each of the $k + (n - k - 1) = (n - 1)$ other treatments. Taking all n rows into account, every treatment is preceded (and succeeded) n times by each of the $(n - 1)$ other treatments. This constitutes proof that the second property holds when the two constructed squares are taken as a whole. Proof that, when n is even, *each* constructed square has this property will be put off until the third property has been dealt with.

In order to facilitate the proof of the third property, we, like Bradley, adopt the concept of 'separation'. Let i be any treatment in a constructed Latin Square, and let j be its immediate predecessor for a given individual. Then let $i - j$ be the difference between the two treatments. If the difference is positive, the 'separation' is equal to that difference. If the difference is negative, the 'separation' is equal to n plus the difference. This concept of separation can be most easily visualized by placing the numbered treatments at equal intervals around a circle in clockwise sequence from 1 to n. The separation between i and $j \neq i$ is the number of intervals passed in moving clockwise from j to i. Note that the separation between any i and every other possible preceding treatment, $j \neq i$, runs from 1 to $n - 1$ in one-to-one correspondence with all treatments other than i. Thus, a design in which, for every treatment, the separations, 1 through $(n - 1)$, appear with equal frequency, is balanced with respect to immediate predecessors.

Now, the cyclic nature of columns has been undisturbed in the construction of the two squares. Consequently, the sequence of separations between successive treatments is the same for every individual row in a given square. We may therefore confine our attention to the sequence of separations in the first row, since the same sequence will apply to every row.

For n odd, treatments in the first row of the left square appear in the sequence, $1, n, 2, n - 1, \cdots, (n + 3)/2, (n + 1)/2$. Then the sequence of differences is alternating in sign, $+(n - 1), -(n - 2), +(n - 3), \cdots, -(1)$. The sequence in the right square is reversed, with changed signs, $+(1), -(2), +(3), \cdots, -(n - 1)$. That is to say, a full set of positive integers from 1 to $n - 1$ and a full set of negative integers from -1 to $-(n - 1)$ appear as differences in the first row of the two squares taken together. When the negative integers are replaced by the corresponding separations, it is seen that in the first (or any) row of the two squares, each separation from 1 to $(n - 1)$ appears exactly twice. Since this is true for every row, and since each treatment appears once in each column, every separation from 1 to $(n - 1)$ occurs exactly twice for each treatment. This establishes that the third property holds for n odd.

For n even, the sequence of differences in the left square is, $+(n - 1), -(n - 2), +(n - 3), \cdots, -(2), +(1)$. Again, this alternating sequence is reversed, with changed signs, in the right square, $-(1), +(2), \cdots, -(n - 3), +(n - 2), -(n - 1)$. Replacing negative differences by the corresponding separations, we get, $(n - 1), 2, (n - 3), \cdots, (n - 2), 1$, *for the left square as well as for the right square.*

This sequence contains all the odd integers in descending sequence from $(n - 1)$ to 1, interlaced with all the even integers from 2 to $(n - 2)$ in ascending sequence. Thus, in the first (or any) row of either square, each separation from 1 to $(n - 1)$ appears exactly once. By the same reasoning as for the case when n is odd, this establishes that the third property holds for each constructed Latin Square when n is even.

It can now be shown that the second property holds for *each* square when n is even. It has already been found that the sequence of separations in the left square is identical with that in the right when n is even. The beginning treatment, together with the sequence of separations, uniquely determines the order of treatment in any row of either square, by reason of one-to-one correspondence of treatments with separations. In the left square, there is just one row which begins with a given treatment, and the right square contains just one row which begins with the same treatment. Consequently, every row in the left square is identical to one and only one row in the right square. Then the rows of the right square could be permuted to arrive at a square which is identical with the left square. These two identical squares are balanced when taken together, if and only if each is balanced. Permutation of rows does not disturb the balance of the right square, and since the two taken together have already been shown to be balanced, each must be balanced with respect to the set of all predecessors.

Graeco-Latin Square Construction.

A slight modification in the procedure for construction, when n is odd, produces a Graeco-Latin Square. This modification was first noted by our colleague, John E. Dowd.

For n odd, proceed in the same way as steps (a), (b) and (c).

Step (d). In the resulting $n \times 2n$ figure, replace treatment numbers appearing in the odd numbered columns by Latin letters and the treatments in even numbered columns by Greek letters.

Now consider all the n^2 mutually exclusive pairs of Latin and Greek letters in adjacent columns. It will be found that the pairs appear in the vertical and horizontal order required to form an $n \times n$ Graeco-Latin Square.

REFERENCES

[1] Williams, E. J. [1949]. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Jour. Sci. Res.*, Series A, V2, 149–68.

[2] Raymond, M. J., Lucas, C. J., Beesley, M. L., O'Connell, B. A., and Roberts, J. A. F. [1957]. A trial of five tranquillizing drugs in psychoneurosis. *Brit. Med. Jour.*, 63–66.

[3] Patterson, H. D. [1952]. The construction of balanced designs for experiments involving sequences of treatments. *Biometrika 39*, 32–48.

[4] Bradley, J. V. [1958]. Complete counterbalancing of immediate sequential effects in a latin square design. *Jour. Amer. Stat. Assoc. 53*, 525–8.

[5] Scheffé, Henry. [1952]. An analysis of variance for paired comparisons. *Jour. Amer. Stat. Assoc. 47*, 381–400.

[6] Cochran, W. G., and Cox, Gertrude. [1957]. *Experimental Designs.* 133–42, second edition, John Wiley & Sons, Inc., N. Y.

[7] Meier, P., Free, S. M., and Jackson, G. L. [1958]. Reconsideration of Methodology in studies of pain relief. *Biometrics 14*, 330–42.

ON THE STATISTICAL THEORY OF A ROVING CREEL CENSUS OF FISHERMEN[1]

D. S. Robson
Cornell University, Ithaca, N. Y., U.S.A

SUMMARY

In order to estimate the day's total catch from a fishery an enumerator roves through the fishing area interviewing fishermen as he encounters them to determine the number n of fish caught and the time t expended. The interviewer is assumed to (i) start his trip at a randomly chosen point along a well defined route which completely covers the fishery, (ii) choose his initial direction at random from the two alternatives, and (iii) travel at a constant rate of c circuits per day. If the catch rate n/t at time of interview is an unbiased estimator of a fisherman's catch rate for his completed trip and if the fishermen's movements relative to the interviewer's path never exceed the interviewer's rate c, then rn/ct, summed over all interviews, is an unbiased estimator of the day's total catch. The unit of time is one day, r is the number of times the fisherman was interviewed, and n/t is the catch rate at the r'th interview.

Unbiasedness of n/t implies that the waiting times to first catch and from first to second catch are identically distributed chance variables, and that all waiting times between successive catches have the same expected value. If waiting times are independent, then unbiasedness implies that fishing is a Poisson process.

INTRODUCTION

Fishing, as every fisherman knows, is a chance process. Skill improves the chances, but a multitude of unknown factors governing fish behavior remain to confound even the most experienced fisherman, and for most of us catching a fish is still largely a matter of chance. One of the basic factors controlling fish catches, of course, is population size; in turn, however, fish population size may itself be strongly influenced through the efforts of the more gifted and more fortunate fishermen. In order to maximize the fishermen's chances insofar as they are influenced by this factor, the fishery manager attempts to maintain the population size and composition at an optimum level

[1] Research supported under a grant by the National Science Foundation while the author was a visiting research associate in the Department of Mathematics, University of Washington.

through such practices as stocking fish and regulating the catch. These management decisions must be based upon objective information regarding the number, size, and age composition of the fishermen's catch, and one of the more important field techniques which have evolved for obtaining such information is the so-called creel census of fishermen.

In the creel census the fishery manager or his representative makes direct observations on the fishing process, interviewing fishermen in action to determine the kinds and numbers of fish taken and the rates at which they are caught. Ordinarily, the fishermen so interviewed represent only a sample of the fishermen present, so that the creel census is, in fact, a sample census. Moreover, in many types of creel census only information on incomplete fishing trips is obtained; that is, the fisherman is interviewed while fishing, providing information on his fishing trip up to the time of interview, and his fortunes after the interview remain unknown to the fishery manager. Such is the case in the type known as the roving creel census, in which the enumerator moves through the fishing area interviewing fishermen as he encounters them, and it is this commonly employed method of sampling and its associated methods of estimation which shall be examined in some detail here.

Estimation of the total day's catch from the fishery on the basis of the roving interviewer's data appears to present a unique combination of statistical problems in the theory of sampling and estimation. Some distinctive features of the roving creel census are (i) the open end to the sample—the number of interviews in the sample is not predetermined but depends, rather, upon the number and distribution of fishermen present, (ii) the sample of fishermen obtained by following some rational route through the fishery constitutes a systematic rather than a random sample, (iii) the probability of interviewing any given fisherman depends in some manner upon how long he fishes, and (iv) only incomplete information is obtained for any one fisherman.

In examining this problem we shall first specify a well defined, roving sampling procedure and then treat the estimation problem under the simplifying assumption that catch rate at the time of interview is an unbiased estimator of that fisherman's catch rate for his completed trip. Later, we consider the implications of this assumption as it relates to the nature of the fishing process.

THE SAMPLING AND ESTIMATION PROCEDURE

A specific description of the procedure followed by a roving interviewer probably would not apply in all detail to any single creel census ever conducted, since each fishery presents circumstances peculiar unto itself. The sampling process to be described here is necessarily a specific idealization of the general plan of roving through the fishery, interviewing fishermen as they are encountered; as always, the idealized plan is not actually attainable in practice, but can be approximated to a reasonable degree.

We shall assume that some systematic route which gives complete coverage is plotted through the fishing area, that the interviewer starts his trip at the beginning of the fishing day from a randomly chosen point of departure along this route, that he chooses at random one of the two alternative directions to travel and then proceeds at a constant rate of travel until the end of the day.

The line denoting the route of the interviewer effectively reduces a fishing area in two dimensions to a line in one dimension, and since the route is closed—that is, a complete coverage of the fishing area will bring the interviewer back to his starting point—then the line may be represented conceptually as a circle. A fisherman's location then corresponds to a point on the circumference of the circle, determined by the point on the interviewer's route at which he would pass that fisherman's location in the fishing area. The dimension of time may be introduced by letting the radius of the circle represent the length of the fishing day. In this way a fisherman's location in both time and space can be represented by a point within the circle; his location in the fishery determines the radius vector upon which he lies and the time of day determines a point on this radius vector. It is convenient here to regard the time axis as extending toward the center of the circle, so that the time is 1 (end of the day) at the center and 0 (beginning of the day) on the circumference. For example, if a fisherman is stationary, then his entire trip can be plotted as a segment of a radius vector as shown in Figure 1a; the particular radius vector is determined by his (fixed) location and the segment extends inward from the time he starts fishing to the time he stops.

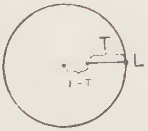


FIGURE 1a

The Mapping of a Stationary Fisherman's Trip Which Starts at the Beginning of the Day and Continues Until Time T at the Location L.

The interviewer's trip, in this framework, then becomes a regular spiral extending from his randomly chosen starting point on the circumference to the center point of the circle. The regularity of the spiral is a direct consequence of the assumption of a constant rate of travel. This is illustrated in figure 1b where the interviewer's rate of travel is arbitrarily taken to be one complete circuit of the fishery per day, and the direction of travel is arbitrarily taken clockwise.

In the particular combination of circumstances described in Figure 1b, the interviewer's trip does not intersect with the fisherman's trip, for the fisherman had already left by the time the interviewer reached that location. Had the interviewer, traveling in this same direction, chosen his starting point anywhere on the arc A shown in Figure 1c then he would have encountered this fisherman, or traveling in the opposite direction and starting anywhere on arc B would have led to an encounter. Since the probability distribution for the starting point is uniform on the circumference of the circle, the probability that it will fall on arc A (or B) is simply the length of the arc expressed as a fraction of the entire circumference. Clearly, this relative length of A (and of B) is simply T, the length of the fisherman's stay; for in order to reach the fisherman's location L at exactly the time T when he stopped fishing, the interviewer would have had to start his trip at a point S just far enough in back of L so as to reach L by traveling for

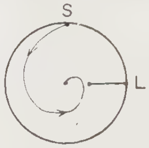


FIGURE 1b

THE MAPPING OF AN INTERVIEWER TRIP WHICH STARTS AT THE POINT S AND MOVES IN A CLOCKWISE DIRECTION AT THE RATE OF ONE CIRCUIT PER DAY

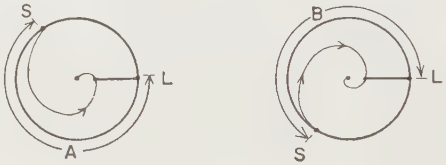


FIGURE 1c

THE RANGE OF INTERVIEWER STARTING POINTS WHICH LEAD TO AN ENCOUNTER WITH THE FISHERMAN AT L, ARC A FOR CLOCKWISE TRIPS, ARC B FOR COUNTER-CLOCKWISE.

exactly a time T, hence covering a T'th of the circumference. The probability that this fisherman would be interviewed is therefore

$$\begin{aligned} P(\text{interview}) &= P(\text{ clockwise travel }) \cdot P(\text{ interview } | \text{ clockwise travel }) \\ &\quad + P(\text{counterclockwise}) \cdot P(\text{interview } | \text{ counterclockwise}) \\ &= \tfrac{1}{2}T + \tfrac{1}{2}T. \end{aligned}$$

Thus, if the interviewer's rate of travel is one circuit per day, then a stationary fisherman's probability of interview is equal to the fraction of a day that he fishes.

If the interviewer's rate of travel is not 1, then this result no longer holds. It is obvious, for example, that, if the interviewer makes 2 complete circuits per day, then every fisherman who fishes for more than half a day is certain to be interviewed at least once, and may be contacted twice. Figure 2a, again employing a stationary fisherman, illustrates this situation and indicates the ranges of starting points which would result in 1 and 2 contacts between interviewer and fisherman. In order for a single contact to occur here, the interviewer must pass the fisherman's location for the second time between time T and time 1. Since he is traveling at the rate of 2 circuits per day, this means that the range of starting points which will accomplish this is of relative length $2(1 - T)$. The remaining range of starting points, of relative length $1 - 2(1 - T)$, will result in 2 contacts between interviewer and fisherman. Figure 2b illustrates the case of a fisherman whose trip length is less than a half day, and indicates the range of starting points which will produce 0 and 1 contact. In order for a single contact to be made the interviewer must pass the fisherman's location between time 0 and time T, and the probability of this occurring is $2T$. We have here ignored the feature of randomized direction of

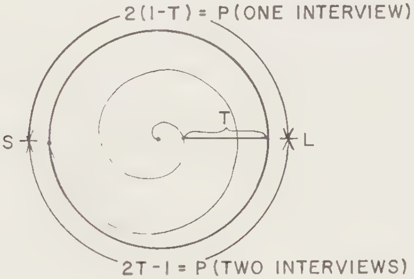


FIGURE 2a

A Mapping of a Stationary Fisherman's Trip of Length $T > 1/2$ and an Interviewer's Trip at a Rate of $c = 2$ Complete Circuits Per Day.

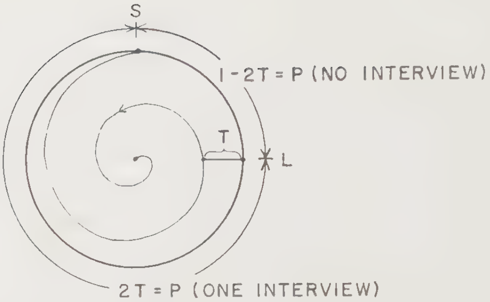


FIGURE 2b

A Mapping of Interviewer and Stationary Fisherman for the Case $c = 2$ and $T < 1/c$.

travel because of the trivial role it plays in the case of stationary fishermen.

In the more general case where the interviewer makes c complete circuits of the fishery in a day then a stationary fisherman whose trip length is T will be interviewed either $[cT]$ or $[cT] + 1$ times, where $[cT]$ denotes the largest integer contained in cT. By the same argument employed earlier, the probability of exactly $[cT]$ interviews is $[cT] + 1 - cT$, and the probability of $[cT] + 1$ interviews is then $cT - [cT]$. The expected number of interviews of this stationary fisherman is therefore

$$[cT]([cT] + 1 - cT) + ([cT] + 1)(cT - [cT]) = cT$$

for any constant travel rate $c > 0$.

If the fishermen themselves are moving about in the fishery then the interviewing process becomes somewhat more complicated analytically, and apparently unmanageable from the viewpoint of estimation. We first observe that if the fisherman's rate of movement relative to the interviewer's path never exceeds the interviewer's rate c then the expected number of interviews remains at cT. To demonstrate this we have exhibited in Figure 3a the path of a slow-moving fisherman,

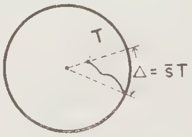
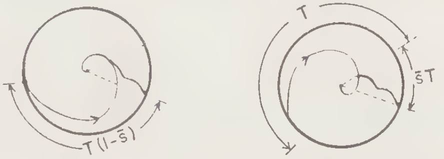


FIGURE 3a

Illustration of a Path Taken by a Moving Fisherman Who Covers a Fraction Δ of the Fishery During a Fraction T of the Day.

traveling in a clockwise direction, who covers a relative distance Δ in a time T, thus moving at an average rate of $\bar{s} = \Delta/T$. As seen in Figure 3b, where the interviewer's rate is taken to be $c = 1$, the range



clockwise route counterwise route

FIGURE 3b

Illustration Showing the Range of Starting Points of Interviewer Trips Which Will Intersect the Moving Fisherman in 3a.

of interviewer starting points which lead to an interview is of relative length $T(1 - \bar{s})$ for clockwise interviewer trips and relative length $T(1 + \bar{s})$ for counterclockwise trips. Thus, the probability of an interview is $T(1 - \bar{s})/2 + T(1 - \bar{s})/2 = T$. More generally, we see that if, for an arbitrary interviewer rate c, the fisherman's rate s is uniformly less than c, then the probability distribution of the number r of interviews for clockwise trips is

$$\begin{aligned} P(r = [T(c - \bar{s})]) &= [T(c - \bar{s})] + 1 - T(c - \bar{s}) \\ &= 1 - P(r = [T(c - \bar{s})] + 1), \end{aligned}$$

and for counterclockwise trips is

$$\begin{aligned} P(r = [T(c + \bar{s})]) &= [T(c + \bar{s})] + 1 - T(c + \bar{s}) \\ &= 1 - P(r = [T(c + \bar{s})] + 1), \end{aligned}$$

and, again, the expected number of interviews is

$$\begin{aligned} \mathcal{E}(r) &= \tfrac{1}{2}\mathcal{E}(r \mid \text{clockwise trip}) + \tfrac{1}{2}\mathcal{E}(r \mid \text{counterclockwise}) \\ &= \tfrac{1}{2}T(c - \bar{s}) + \tfrac{1}{2}T(c + \bar{s}) \\ &= cT. \end{aligned}$$

As soon as the fisherman's rate s exceeds the interviewer's rate c, the roles of the two paths effectively become reversed in so far as they determine the number of interviews. In effect, the hunter becomes the hunted, and an interview occurs now if the fisherman overtakes the interviewer. Consequently, if a fisherman travels at a constant rate $s > c$ for a time T then the expected number of times he will be interviewed during this period is sT rather than cT. If during a period of length T the fisherman's rate $s > c$ is not constant but is s_1 for a time τ_1 then s_2 for a time τ_2 , $\cdots$, then s_k for a time τ_k , $\tau_1 + \cdots + \tau_k = T$, then his expected number of interviews during this period is $\bar{s}T$, where

$$\bar{s} = \frac{1}{T} \sum_{i=1}^{k} s_i \tau_i \ .$$

It then follows that, if s varies continuously, $s = s(\tau) > c$, in the interval $0 < \tau < T$, the expected number of interviews is

$$\mathcal{E}(r) = \bar{s}T = \int_0^T s(\tau)\, d\tau.$$

For a completely arbitrary type of fishing trip of duration T, the fisherman's rate of movement $s(\tau)$ may exceed c part of the time, and part of the time not, so in general

$$\mathcal{E}(r) = c \int_{\mathfrak{J}-\mathfrak{J}_c} d\tau + \int_{\mathfrak{J}_c} s(\tau)\, d\tau,$$

where

$$\mathfrak{J} - \mathfrak{J}_c = \{\tau \mid 0 \le \tau \le T, s(\tau) \le c\},$$
$$\mathfrak{J}_c = \{\tau \mid 0 \le \tau \le T, s(\tau) > c\}.$$

Since there will be only finitely many discontinuities in $s(\tau)$, we may simply write

$$\mathcal{E}(r) = cT + (\bar{s} - c)T_c \ ,$$

where T_c is the Riemann measure of $\mathfrak{J}_c$, or the total length of time that $s(\tau) > c$, and $\bar{s}$ is the average of $s(\tau)$ over $\mathfrak{J}_c$,

$$\bar{s} = \frac{1}{T_c} \int_{\mathfrak{J}_c} s(\tau)\, d\tau.$$

The two components of $\mathcal{E}(r)$, cT and $(\bar{s} - c)T_c$, are not individually estimable on the basis of the interviewer's data—unless, perhaps, he also makes some quantitative measure of the fisherman's rate of movement when he is approached for an interview. For this reason we shall, from this point onward, assume that no fisherman's rate of movement ever exceeds that of the interviewer ($T_c = 0$), and this is most easily accomplished by imposing the restriction that the fisherman's time in motion is not counted as fishing time, nor is he interviewed while in motion.

At this point it is worth noting that while the arguments so far have been directed at the case of a single interviewer traveling at the constant rate of c circuits per day, all arguments apply just as well to the case of k interviewers equally spaced along the route and traveling at the constant rate of c/k. The combined data of k such interviewers is equivalent in every way to the data of a single interviewer traveling k times as fast; that is, traveling at the rate c.

The preceding results apply to each fisherman who is present in the fishery sometime during the course of the day; consequently, if there are M fishermen present with trip lengths $T_1, \cdots, T_M$, then the expected number of interviews is $c(T_1 + \cdots + T_M)$. If $c < 1$, then this is also the expected number of different fishermen interviewed; otherwise, the expected number of interviews exceeds the expected number of different fishermen contacted by an amount $c \sum^{+} (T_i - 1/c)$, where the sum $\sum^{+}$ extends over all fishermen whose effort T_i exceeds $1/c$.

In his interviews the enumerator determines the number n of fish caught and the amount of effort (time) t expended up to the time of interview, thus enabling him to estimate the catch rate for each interviewed fisherman. Conventional methods of estimating the day's total catch from the fishery by a roving creel census are based upon the assumption that this catch rate n/t at the time of interview is an unbiased estimator of the catch rate N/T for the completed trip. We shall examine the implications of this assumption in a later section; for the moment, however, we will accept it. If the fisherman is interviewed r times during his trip, then r such unbiased estimates of N/T are avaliable; since the information is cumulative, however, the last interview contains all of the information of those preceding it, so the last catch rate would be used to estimate that fisherman's N/T. Using the established fact that the expected value of r, the number of interviews, is cT, we shall then see that rn/ct is an unbiased estimate of N, and summing this over all m fishermen interviewed then gives an unbiased estimate of the total number of fish caught by all M fishermen during the day.

A simple numerical example shown in Figure 4 illustrates this point. Here the interviewer's rate of travel is $c = \frac{3}{2}$ circuits per day; two stationary fishermen are present during the day, one starting at the beginning of the day and fishing a fraction $T_1 = \frac{7}{8}$ of the day and the second, who is $\frac{3}{8}$ of a circuit behind the first on the interviewer's (clockwise) route, starts fishing at time $\frac{3}{8}$ and fishes until time $\frac{1}{2}$ ($T_2 = \frac{1}{8}$). The first fisherman will then be interviewed either once or twice, $r_1 = 1$ or 2, and $r_2 = 0$ or 1, giving four different possible outcomes for r_1 and r_2. The ranges of interviewer starting points which produce each of these four outcomes are shown in Figure 4 along with their relative lengths,

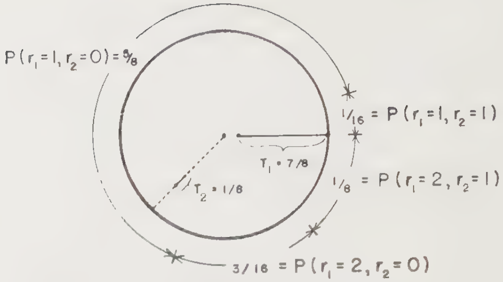


FIGURE 4

A NUMERICAL EXAMPLE WITH TWO STATIONARY FISHERMEN SHOWING THE PROBABILITY DISTRIBUTION OF NUMBER OF INTERVIEWS WHEN THE INTERVIEWER'S TRAVEL RATE IS $c = 3/2$.

or probability measures. The expected value of $[(r_1N_1/T_1) + (r_2N_2/T_2)]/c$ in this example is then

$$\frac{2}{3}\left[\frac{1}{16}\left(\frac{N_1}{T_1} + \frac{N_2}{T_2}\right) + \frac{1}{8}\left(\frac{2N_1}{T_1} + \frac{N_2}{T_2}\right) + \frac{3}{16}\left(\frac{2N_1}{T_1}\right) + \frac{5}{8}\left(\frac{N_1}{T_1}\right)\right]$$

$$= \frac{2}{3}\left[\frac{21}{16}\left(\frac{N_1}{T_1}\right) + \frac{3}{16}\left(\frac{N_2}{T_2}\right)\right]$$

$$= \frac{2}{3}\left[\frac{21}{16}\left(\frac{8}{7}\right)N_1 + \frac{3}{16}\left(\frac{8}{1}\right)N_2\right] = N_1 + N_2 \ .$$

The sampling error of an estimate of this type may be regarded as a sum of two uncorrelated components,

$$\frac{1}{c}\sum_{(m)}\frac{rn}{t} - \sum_{(M)} N = \left[\frac{1}{c}\sum_{(m)}\frac{rN}{T} - \sum_{(M)} N\right] + \left[\frac{1}{c}\sum_{(m)} r\left(\frac{n}{t} - \frac{N}{T}\right)\right],$$

where $\sum_{(m)}$ is a sum over the m fishermen sampled and $\sum_{(M)}$ is a sum over the M fishermen present, the first due to the sampling of fishermen and the second due to the sampling of the fishing process of the selected fisherman—that is, due to the incompleteness of the fishing trip at the time of interview. These components are statistically uncorrelated because of our assumption that for any given t, $0 < t < T$, the observed catch rate n/t is an unbiased estimator of N/T. Unfortunately, however, the sampling variance and its corresponding components are in general not estimable from the data obtained in a roving census because of the systematic nature of the sample and unequal, unknown probabilities of selection.

We turn now to an examination of our simplifying assumption that catch rate at time of interview is an unbiased estimator of the catch rate for the completed trip. This assumption clearly relates to the nature of the stochastic fishing process, and in the next section we will study this relationship in some detail.

IMPLICATIONS OF THE ASSUMPTION OF AN UNBIASED CATCH RATE

For this discussion we again restrict our attention to a single fisherman on a given day and on a fishing trip of given duration T. Since fishing is a chance process, the number N_T of fish which will be caught during this period is a chance variable. The basic chance variable of the fishing process, however, is the amount of time required to catch a fish, or the waiting time between successive catches. The number N_T is simply the number of successful waiting times contained in an interval of length T; hence the probability distribution of N_T depends basically upon the distribution of waiting times.

Our fisherman is being subjected to a sequence of waiting times $w_1, w_2, w_3, \cdots$ between successive catches, and, since he terminates the process after a period of length T, then the number N_T of fish caught will be that number K for which $w_1 + w_2 + \cdots + w_K \leq$ T $< w_1 + w_2 + \cdots + w_K + w_{K+1}$. Thus, if W_K denotes the sum of the first K waiting times, then the probability distribution of N_T is

$$P(N_T = K) = P(W_K \leq T < W_{K+1}).$$

The waiting times $w_1, w_2, w_3, \cdots$ are non-negative chance variables, so the event $N_T = K$ is also defined by excluding the event $W_K > T$ from the event $W_{K+1} > T$; consequently, an equivalent expression is

$$P(N_T = K) = P(W_{K+1} > T) - P(W_K > T),$$

or, in the classical form for waiting time problems,

$$P(N_T = K) = P(W_K \leq T) - P(W_{K+1} \leq T). \tag{1}$$

The catch n_t on hand at the time of interview of our fisherman is also a chance variable whose distribution then depends on t, T and N_T. To compute this distribution we note that the joint event $n_t > k$ and $N_T = K$ occurs if and only if $W_{k+1} \leq t$ and $W_K \leq T < W_{K+1}$; thus, the conditional probability of $n_t > k$, given t, T and N_T, is

$$P(n_t > k \mid t, T, K) = \frac{P(W_{k+1} \leq t, W_K \leq T < W_{K+1})}{P(W_K \leq T < W_{K+1})}$$

$$= \frac{P(W_{k+1} \leq t, W_K \leq T) - P(W_{k+1} \leq t, W_{K+1} \leq T)}{P(W_K \leq T) - P(W_{K+1} \leq T)}.$$

From this we obtain the distribution of n_t by subtraction,

$$P(n_t = k \mid t, T, N_T) = P(n_t > k - 1 \mid t, T, N_T) - P(n_t > k \mid t, T, N_T).$$

This derivation of the distribution of N_T and n_t implicitly assumes that the duration T of the fishing trip is an arbitrary constant, determined independently of the outcomes $w_1, w_2, \cdots$ of the waiting process. While this is possibly true of the fishing process, as when the fisherman decides in advance just how long he will fish, we must also acknowledge the existence of other possible stopping rules, including those which are a sequential function of the waiting times. When we impose the restriction of unbiasedness, however, we require that all admissible stopping rules exhibit the same dependence on waiting times, and, since the rule in which T is an arbitrary constant must be included among the possible stopping rules, then all admissible rules must be independent of waiting times. This is clearly seen in the case $N_T = 1$, where the condition of unbiasedness becomes

$$P(n_t > 0 \mid t, T, N_T = 1) = P(w_1 \leq t \mid t, T, N_T = 1) = \frac{t}{T}.$$

Here we see that the conditional distribution of w_1, given that fishing stops at time T with $N_T = 1$ fish in the creel, must be the uniform distribution from 0 to T, regardless of the stopping rule. Now if the process satisfies this condition when T is an arbitrary constant then every other stopping rule for which this condition is satisfied can contain no more information concerning w_1 than does the rule in which T is arbitrary. We shall continue to assume, therefore, that T is simply a preassigned constant for each fisherman.

The preceding arguments also assume that the restriction of unbiasedness applies conditionally for all t, and we shall now demonstrate that this must, in fact, be true if we require that unbiasedness hold for an arbitrary travel rate c. To see this, let

$$h_{T,K}(t) = \mathcal{E}\left(\frac{n_t}{t} \,\middle|\, t, T, N_T = K\right),$$

so that the condition of unbiasedness becomes

$$\frac{K}{T} = \int_0^T h_{T,K}(t)\, dP_T(t).$$

For a stationary fisherman the distribution of t, the time of the last interview, is uniform from $T - (1/c)$ to T when $T \geq (1/c)$. Since the choice of c, the interviewer's travel rate, is arbitrary relative to the time T fished by this particular fisherman, then for all $c > (1/T)$

$$\frac{K}{T} = c \int_{T-(1/c)}^{T} h_{T,K}(t)\, dt,$$

or, letting

$$H_{T,K}(t) = \int h_{T,K}(t)\, dt,$$

then

$$H_{T,K}(T) - H_{T,K}\left(T - \frac{1}{c}\right) = \frac{K}{T}\left(\frac{1}{c}\right).$$

Letting $x = T - (1/c)$ we then have

$$H_{T,K}(x) = H_{T,K}(T) - \frac{K}{T}(T - x),$$

and upon differentiating with respect to x,

$$h_{T,K}(x) = \frac{K}{T},$$

for $0 < x < T$. This justifies our earlier assertion that our unbiased condition on n/t implies that

$$\mathcal{E}\left(\frac{rn}{ct}\right) = N,$$

for now we may write

$$\mathcal{E}\left(\frac{rn}{ct}\right) = \mathcal{E}\left\{\frac{r}{ct}\,\mathcal{E}(n \mid r, t)\right\} = \frac{N}{cT}\,\mathcal{E}(r) = N.$$

Returning to the distribution of n_t, we may express the unbiasedness condition as

$$\sum_{k=0}^{K} kP(n_t = k \mid t, T, N_T = K) \equiv \frac{tK}{T},$$

or, equivalently,

$$\sum_{k=0}^{K-1} P(n_t > k \mid t, T, N_T = K) \equiv \frac{tK}{T}.$$

In terms of the distribution of waiting times this identity then becomes

$$\sum_{k=1}^{K} [P(W_k \leq t, W_K \leq T) - P(W_k \leq t, W_{K+1} \leq T)] \equiv \frac{tK}{T}[P(W_K \leq T) - P(W_{K+1} \leq T)] \tag{2}$$

which must hold for all t and T, $0 < t < T < \infty$ and for every positive integer K (for $K = 0$ the identity is trivial).

The identity (2) imposes severe restrictions on the nature of the fishing process as it is described by the distribution of waiting times. For the special case $K = 1$; i.e., the fisherman's total catch is one fish, this identity reveals several facts. If the cumulative distribution function of w_1, the time required to catch the first fish, is denoted by $F(w_1)$,

$$F(x) \equiv P(w_1 \leq x),$$

and, if

$$G(y \mid x) \equiv P(w_2 \leq y \mid w_1 = x),$$

then

$$H(z) = \int_0^z G(z - x \mid x)\, dF(x) = P(W_2 < z)$$

and (2) becomes, for $K = 1$,

$$F(t) - \int_0^t G(T - x \mid x)\, dF(x) \equiv \frac{t}{T}\,[F(T) - H(T)].$$

Differentiating both sides with respect to t we obtain a new identity

$$f(t) - G(T - t \mid t)f(t) \equiv \frac{1}{T}\,[F(T) - H(T)]. \tag{3}$$

Now setting $t = T$ we find that

$$f(T) = \frac{1}{T}\,[F(T) - H(T)], \tag{4}$$

so

$$P(N_T = 1) = Tf(T). \tag{5}$$

Differentiating (4) with respect to T gives

$$h(T) = -Tf'(T), \tag{6}$$

that is, the density function of the sum of the first two waiting times at the point T must be equal to T times the negative derivative of f at T. This, in turn, implies that *the density function f of the first waiting time is monotonically decreasing*; that is, if $0 < w < w'$ then $f(w) > f(w')$.

Combining (3) and (4) we have

$$f(t)[1 - G(T - t \mid t)] = f(T) \tag{7}$$

and now differentiating with respect to T we find

$$f(t)g(T - t \mid t) = -f'(T),$$

or

$$f(w_1)g(w_2 \mid w_1) = -f'(w_1 + w_2). \tag{8}$$

Thus, the joint distribution of the first and second waiting times, w_1 and w_2, must be equal to the negative derivative of the density function f, evaluated at the point $w_1 + w_2$. Furthermore, upon integrating w_1 out of this joint density function we find the marginal density function of w_2 to be

$$\int_0^\infty f(w_1)g(w_2 \mid w_1)\, dw_1 = -\int_0^\infty f'(w_1 + w_2)\, dw_1 = f(w_2),$$

that is, w_1 *and* w_2 *must be identically distributed.* Note that for fixed w_1 the mean value of w_2 must be

$$\mathcal{E}(w_2 \mid w_1) = [1 - F(w_1)]/f(w_1).$$

If w_1 and w_2 are independent as well as being identically distributed, then we see, upon returning to (7) and differentiating with respect to t, that

$$f'(t)[1 - F(T - t)] + f(t)f(T - t) = 0,$$

or, putting $t = T$,

$$f'(T)/f(T) = -f(0) = -\theta \quad \text{(say)}$$

so that the distribution $f(T)$ of the first and second waiting times must be the exponential

$$f(T) = \theta e^{-\theta T}.$$

From (8) we also see that the converse is true; that is, if w_1 and w_2 are identically, exponentially distributed then they must be independent if our condition of unbiasedness is to be fulfilled.

Most of these arguments and conclusions apply in general to the original identity (2) holding for arbitrary K. In general, (4) becomes

$$dP(W_K \leq T)/dT = \frac{K}{T}\,[P(W_K \leq T) - P(W_{K+1} \leq T)],$$

so that (5) becomes

$$P(N_T = K) = \frac{T}{K}\,\frac{dP(W_k \leq T)}{dT}.$$

The relation (6) becomes, in general,

$$\frac{dP(W_{K+1} \leq T)}{dT} = \frac{(-1)^K T^K f^{(K)}(T)}{K!}, \tag{9}$$

where $f^{(K)}(T)$ is the K'th derivative of $f(T)$. Combining these last two equations we find that

$$P(N_T = K) = (-1)^{K-1} \frac{T^K}{K!} f^{(K-1)}(T) \tag{10}$$

for $K > 0$, and

$$P(N_T = 0) = 1 - F(T).$$

From (10) it is seen that the generating function $Q_T(s)$ of the distribution of N_T takes the form

$$\begin{aligned} Q_T(s) &= \sum_{K=0}^{\infty} s^K P(N_T = K) \\ &= 1 - F(T) + sTf(T) - \frac{(sT)^2}{2!} f'(T) + \frac{(sT)^3}{3!} f''(T) - \cdots \end{aligned}$$

which is the Taylor series expansion of

$$Q_T(s) = 1 - F(T - Ts) \tag{11}$$

about the point T. The factorial moments of N_T are therefore

$$\mathcal{E} N_T(N_T - 1) \cdots (N_T - r + 1) = (-1)^{r-1} T^r f^{(r-1)}(0).$$

In particular, the mean value of N_T is

$$\mu_T = Tf(0),$$

and the variance is

$$\sigma_T^2 = \mu_T(1 - \mu_T) - T^2 f'(0).$$

The earlier argument for $K = 1$ which led to (8) and the conclusion that

$$dP(w_2 \leq x) = dP(w_1 \leq x),$$

now gives for $K = 2$

$$dP(w_3 \leq x) + dP(w_2 + w_3 \leq x) = dP(w_1 \leq x) + dP(w_1 + w_2 \leq x),$$

and for $K = 3$

$$\begin{aligned} &dP(w_4 \leq x) + dP(w_3 + w_4 \leq x) + dP(w_2 + w_3 + w_4 \leq x) \\ &\qquad = dP(w_1 \leq x) + dP(w_1 + w_2 \leq x) + dP(w_1 + w_2 + w_3 \leq x), \end{aligned}$$

or, in general,

$$\begin{aligned}\sum_{k=1}^{K} dP(W_{K+1} - W_k \le x) &= \sum_{k=1}^{K} dP(W_k \le x) \\ &= \sum_{k=1}^{K} (-1)^{k-1} \frac{x^{k-1}}{(k-1)!} f^{(k-1)}(x)\,dx.\end{aligned} \tag{12}$$

The implication for $K + 1$ waiting times is thus not quite as specific as it was for 2 waiting times—namely, that the first two must be identically distributed. However, multiplying the system (12) by x and integrating does show that *all waiting times must have the same expected value.*

The conclusions concerning independence and the exponential distribution do extend to arbitrary K. *If waiting times are mutually independent, then $f(w_1) = f(w_2)$ is exponential*, and from (9) we see that the distribution of the sum W_K is simply the K-fold convolution of f, which implies that *all waiting times are distributed according to f.* As is well known, independent and identically exponentially distributed waiting times generate a Poisson distribution for number of occurrences, and in this case (10) gives us

$$P(N_T = K) = e^{-\theta T} \frac{(\theta T)^K}{K!},$$

and from (11) we now obtain the Poisson generating function

$$1 - F(T(1 - s)) = 1 - [1 - e^{-\theta T(1-s)}] = e^{-\theta T(1-s)}.$$

When fishing is a Poisson process, the conditional distribution of n_t, the number of fish on hand at the time of interview, is binomial with parameters N_T and $p = t/T$, for substituting the exponential waiting time distribution into our formulas we obtain

$$\begin{aligned}P(n_t = k \mid t, T, N_T = K) &= \frac{K!}{T^K}\left[\frac{1}{k!(K - k - 1)!}\int_0^t x^k (T - x)^{K-k-1}\,dx\right. \\ &\left. - \frac{1}{(k-1)!(K-k)!}\int_0^t x^{k-1}(T - x)^{K-k}\,dx\right] = \binom{K}{k}\left(\frac{t}{T}\right)^k \left(1 - \frac{t}{T}\right)^{K-k}.\end{aligned}$$

Thus, for all t, $0 < t \le T$, the catch rate at t is unbiased,

$$\mathcal{E}\left(\frac{n_t}{t} \,\middle|\, t, N_T\right) = \mathcal{E}\left(\frac{n_t}{t} \,\middle|\, N_T\right) = \frac{N_T}{T},$$

and here the conditional variance of the catch rate estimate is

$$\operatorname{var}\left(\frac{n_t}{t} \,\middle|\, t, N_T\right) = \frac{N_T}{T}\left(\frac{1}{t} - \frac{1}{T}\right).$$

The above variance is conditional upon t as well as T and N_T; since we know the conditional distribution of t, the time of the last interview, we can in this case compute the value of the second component of variance of the estimator of $\sum N$—that is, we can now compute

$$\mathcal{E}\left[\frac{1}{c}\sum_{(m)} r\left(\frac{n_t}{t}-\frac{N_T}{T}\right)\right]^2.$$

for fixed T's and N_T's. Assuming that the different fishermen are undergoing independent Poisson processes, we see that

$$\mathcal{E}\left[\frac{1}{c}\sum_{(m)} r\left(\frac{n_t}{t}-\frac{N_T}{T}\right)\right]^2 = \frac{1}{c^2}\,\mathcal{E}\sum_{(m)} r^2\,\frac{N_T}{T}\left(\frac{1}{t}-\frac{1}{T}\right),$$

where

$$P(r=[cT]) = 1 - cT + [cT] = 1 - P(r=[cT]+1),$$

and where the conditional distribution of t given r is uniform. For $r = [cT]$, t must fall in the interval $(T - 1/c, [cT]/c)$ of length $(1/c)P(r = [cT])$ and for $r = [cT] + 1$, t must fall in the interval $([cT]/c, T)$ of length $(1/c)P(r = [cT] + 1)$. From this joint distribution of r and t we find the second component of variance to be

$$\mathcal{E}\left\{\frac{1}{c}\sum_{(m)} r\left(\frac{n_t}{t}-\frac{N_T}{T}\right)\right\}^2 = \sum_{(M)} N_T\left\{\frac{[cT]^2}{cT}\log\frac{[cT]}{cT-1} + \frac{([cT]+1)^2}{cT}\log\frac{cT}{[cT]} - \left(1-\frac{[cT]}{cT}\right)\left(\frac{[cT]+1}{cT}-1\right)-1\right\}.$$

As c gets large then the probability of interview approaches 1 for every fisherman, and the time t converges to T, so this component (as well as the first component) approaches 0.

Examples of other stochastic processes which satisfy our condition (2) of unbiasedness can now be easily constructed by the simple device of mixing Poisson processes. Since unbiasedness holds conditionally for a Poisson process with parameter θT then it must also hold unconditionally when θ is assigned some probability measure $P(\theta)$. In effect, then, the fisherman is assigned a value of θ for his trip, where the θ is selected according to the distribution $P(\theta)$. The density function of the waiting time to first catch is then

$$f(w_1) = \int_0^\infty \theta e^{-\theta w_1}\,dP(\theta).$$

All waiting times will have this same marginal distribution, and their joint distribution will be

$$f(w_1, \cdots, w_k) = \int_0^\infty \theta^k e^{-\theta(w_1+\cdots+w_k)}\, dP(\theta).$$

The covariance between any two will be the same as the covariance of w_1 and w_2, which can be written in general as

$$\operatorname{cov}(w_1, w_2) = \int_0^\infty x(1 - F(x))\, dx - \left\{\int_0^\infty xf(x)\, dx\right\}^2,$$

and which in the present case becomes

$$\operatorname{cov}(w_1, w_2) = \int_0^\infty \frac{dP(\theta)}{\theta^2} - \left[\int_0^\infty \frac{dP(\theta)}{\theta}\right]^2 = \operatorname{var}\left(\frac{1}{\theta}\right).$$

The distribution of N_T is simply

$$P(N_T = K) = \int_0^\infty P(N_T = K \mid \theta)\, dP(\theta)$$
$$= \frac{T^K}{K!}\int_0^\infty \theta^K e^{-T\theta}\, dP(\theta),$$

which, of course, could also be obtained from (10) using this function f. Since $P(n_t \mid t, T, N_T, \theta)$ is binomially distributed independently of θ, then this is also the unconditional distribution,

$$P(n_t \mid t, T, N_T) = \int_0^\infty P(n_t \mid N_T, \theta)\, dP(\theta)$$
$$= \binom{N_T}{n_t}\left(\frac{t}{T}\right)^{n_t}\left(1 - \frac{t}{T}\right)^{N_T - n_t},$$

regardless of the mixing distribution $P(\theta)$.

To take a specific example of this mixing procedure, let $P(\theta)$ be the gamma distribution

$$dP(\theta) = \frac{\beta^\alpha \theta^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\theta}\, d\theta,$$

then

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta} + 1\right)^{-(\alpha+1)},$$

and

$$f^{(K-1)}(x) = (-1)^{K-1}\frac{\Gamma(\alpha + K)}{\beta^K\Gamma(\alpha)}\left(1 + \frac{x}{\beta}\right)^{-(\alpha+K)}.$$

The distribution of N_T, a gamma mixture of Poisson distributions, is from (10) the negative binomial

$$P(N_T = K) = \frac{\Gamma(\alpha + K)}{K!\Gamma(\alpha)} \left(\frac{T}{\beta}\right)^K \left(1 + \frac{T}{\beta}\right)^{-(\alpha+K)}.$$

A negative binomial process derived in this manner is of no particular interest as such, because the basic process which the fisherman follows on any given trip is still Poisson. We may now change our viewpoint, however, and regard this method of derivation as merely a device for showing that fishing can be a negative binomial process and still satisfy our condition that catch rate is unbiased. We stand on weaker ground, of course, if we know that a fishing trip is a negative binomial process, for we cannot then be certain of unbiasedness as we could in the Poisson case. The same would be true of any other distribution we might generate by this device of mixing Poisson processes.

AN EXAMPLE OF BIAS

To illustrate the magnitude of the bias which might arise if the fishing process does not satisfy the conditions of the preceeding section we may consider a simple example where successive waiting times are independent and identically but not exponentially distributed. In particular, if we let waiting time w be proportional to a chi-square variable with an even number of degrees of freedom, say $w = \chi^2_{2\upsilon}/2\beta$, then the computation problem remains relatively simple. For $\upsilon = 1$, of course, we again obtain the exponential distribution of waiting time. For $\upsilon > 1$ the waiting time distribution will no longer be a decreasing function with its maximum at the origin; rather, it will increase to a maximum at the point $w = (\upsilon - 1)/\beta > 0$ and then decrease. We are thus assuming that the wait between two successive catches is more likely to last 10–20 minutes, say, than 0–10 minutes, and that perhaps a wait of 20–30 minutes is still more likely than a 10–20 minute wait, though eventually the probability reaches a maximum and then steadily decreases. This might be the situation, for example, in fishing for some of the more voracious species of game fish which tend to overdisperse themselves, exhibiting a negative or repulsive contagion in their spacial distribution. Catching one of these fish almost automatically precludes the immediate capture of another.

For illustrative purposes, then, let $\upsilon = 2$, so that the distribution of each waiting time w_i is

$$dP(w_i \leq w) = \beta^2 w e^{-\beta w}\, dw.$$

Then if the interviewer's travel rate is $c \leq 1$ it can be easily shown from the preceding sections that for a fixed T

$$\mathcal{E}\left(\frac{rn}{ct}\,\middle|\,N_T = 1\right) = \frac{9 + 5\beta T}{18 + 6\beta T},$$

and

$$\mathcal{E}\left(\frac{rn}{ct}\,\middle|\,N_T = 2\right) = \frac{400 + 12T + 65\beta + 27\beta T}{300 + 60\beta T}.$$

Since $0 < \beta < \infty$, we see that

$$\frac{1}{2} < \mathcal{E}\left(\frac{rn}{ct}\,\middle|\,N_T = 1\right) < \frac{5}{6},$$

and

$$\frac{4}{3} + \frac{T}{25} < \mathcal{E}\left(\frac{rn}{ct}\,\middle|\,N_T = 2\right) < \frac{13}{12T} + \frac{9}{20}.$$

Thus, if a fisherman catches exactly 1 fish on his trip then his expected contribution to the interviewer's estimate of the total catch is somewhere between 1/2 and 5/6 rather than the desired 1. A fisherman who catches 2 fish may be expected to contribute somewhere between 4/3 and ∞ to the estimate, depending upon how long he fishes. If he fishes all day ($T = 1$) to catch the 2 fish then his expected contribution is between 1.373 and 1.533 fish rather than 2; if he fishes only a very short time, say $T = .01$, then his expected contribution to the estimate is somewhere between 1.333 and 108.783, and is more likely to be toward the large end since that fisherman's β-parameter is likely to be large.

Clearly, if a collection of fishermen are present with varying T's and β's, the interviewer's estimate in this case could not be expected to bear any particular resemblance to the actual total catch.

DISCUSSION

The major weakness of the roving creel census of incompleted fishing trips as a technique for estimating total catch is that the bias of estimation depends on the basic nature of the stochastic fishing process, and this, in general, is unknown. We can only speculate as to the nature of the fishing process and whether or not it satisfies the conditions of unbiasedness, but in some circumstances the answer is obvious. If the fish occur in schools and several may be captured whenever a school is encountered, then waiting time to first fish is waiting time to first school, but waiting time from first to second fish may be waiting time between catches within a school, and these two are not identically distributed chance variables. Or the habit of the fisherman may be to visit the more productive locations during the first part of his trip and then spend some time exploring for more. A variety of arguments could be put forward for unequal expected waiting times in violation of our conditions for unbiasedness, and the resulting bias could be of considerable magnitude.

The only sure way of avoiding this problem in a creel census is to use some other technique in place of the roving interviewer, to make the creel census *distribution-free* in the sense of an ordinary sample survey method. One effective, distribution-free technique is to station interviewers at the access points of the fishery to obtain information on completed trips. When no well defined access points exist, interviewers may be assigned randomly chosen area segments of the fishery in which they are to obtain completed-trip information [1]. References to these and other techniques may be found in an extensive bibliography on game kill and creel census procedures prepared by V. Schultz [2].

It should be noted that the implications concerning the fishing process which were deduced from the condition of unbiasedness of the catch rate at time of interview apply to the general problem of predicting the number of events in a stochastic process. If the number n_t of events in a time-continuous process is observed after the process has been in operation for a time t, then in order for n_tT/t to be an unbiased predictor of the number of events N_T which will have occurred at some specified later time T, the process must satisfy the conditions described here.

There are also other applications in which the role of the time dimension is played by some other continuous variable; for example, in a process where objects are randomly drawn until their combined weight first exceeds T pounds and then a sample of these objects is withdrawn until the weight of the sample first exceeds t pounds, $t < T$, then $(n_t + 1)T/t$ will be an unbiased predictor of the number $N_T + 1$ of objects in the T-pound sample only if the weight of individual objects is exponentially distributed. An interesting modification of this problem arises when the exact weights of the sample and/or subsample are known; that is, when the excess weights above T and t pounds, respectively, are also measured. In the analogous fishing-interviewing process it might be possible for the interviewer to determine that time t', $t' \leq t$, at which the n_t'th fish was caught, and use of this type of information would certainly modify the form and properties of the predictor of total catch. Such problems involving additional interview information warrant further investigation.

ACKNOWLEDGMENTS

The author wishes to acknowledge M. D. Grosslein, formerly at Cornell University, for his assistance in formulating the roving creel census problem, and D. G. Chapman, University of Washington, for his helpful comments during the development of the paper.

REFERENCES

[1] Robson, D. S. [1960]. An unbiased sampling and estimation procedure for creel censuses of fishermen. *Biometrics 16*, 261–77.

[2] Schultz, V. [1959]. A contribution toward a bibliography on game kill and creel census procedures. *Misc. Publ. No. 359, Md. Agric. Exp. Sta.*

ALLOCATION OF EXPERIMENTAL UNITS IN SOME ELEMENTARY BALANCED DESIGNS

J. S. Williams
Research Triangle Institute[1]
Durham, North Carolina, U.S.A.

INTRODUCTION

In recent years an increased amount of effort has been devoted to studies of how best to allocate experimental or sampling units for estimating variance components. Concepts which originated with the sample survey workers for estimating means and totals are reviewed in Cochran's text [1953]. Crump [1954] and Gaylor [1960] have extended these concepts to the estimation of variance components in nested and cross-classified designs.

The usual procedure of allocation is to minimize a set of specified variances, often subject to given cost restrictions and a fixed total number of units. This can obscure the fact that the fixed number of units automatically limits the amount of attention which can be devoted to any one component. Also it is superficial to minimize variances subject to cost restrictions, if cost is not a major consideration. Achieving a predetermined balance of variances of the estimates sometimes is more meaningful, but frequently as will be shown, this is not obtainable with a balanced allocation.

The following development illustrates these principles for balanced, two-level nested allocations and two-way cross-classified allocations of normally distributed variables. The assumption of normality usually is justified for applications where variance component analysis is important, such as in genetic and chemical studies.

GENERAL DEVELOPMENT

In the simple two-level, nested analysis of variance, there are n_1 primary (or first level) experimental units, n_2 secondary (or second level) experimental units for each primary unit, and $N = n_1 n_2$ units in all. The two components of variance which can be estimated are σ^2 the within, estimated by the within mean square $SSW/n_1(n_2 - 1) = \hat{\sigma}^2$, and the among $\sigma_a^2 = k\sigma^2$ (say), estimated by the difference between the among and within mean squares,

[1] Part of this paper was written while the author was a graduate student at North Carolina State College.

$$[SSA/(n_1 - 1)] - [SSW/n_1(n_2 - 1)] = n_2\hat{\sigma}_a^2 .$$

The circumstances of the populations to be sampled may dictate that either (*i*) n_1 is predetermined and n_2 is any integer or (*ii*) N is fixed and (n_1 , n_2) is a pair of integers such that $n_1n_2 = N$.

(*i*) For the assumption of normality,

$$\text{Var } [\hat{\sigma}_a^2] = \frac{2}{n_2^2} \frac{n_1(n_2 - 1)(1 + n_2k)^2 + (n_1 - 1)}{n_1(n_1 - 1)(n_2 - 1)} \sigma^4, \tag{1}$$

$$\text{Var } [\hat{\sigma}^2] = \frac{2\sigma^4}{n_1(n_2 - 1)}. \tag{2}$$

In Var $[\hat{\sigma}_a^2]$, the variance of $SSA/(n_1 - 1)$,

$$\frac{2\sigma^4}{(n_1 - 1)} \left[\frac{1}{n_2^2} + \frac{2k}{n_2} + k^2\right],$$

dominates for all but very small values of n_1 and k. For fixed n_1 , the most important term of this variance is k^2, which is determined by the uncontrolled, physical properties of the population sampled. When k is not small, large values of n_2 can have little effect on Var $[\hat{\sigma}_a^2]$ while they provide a more reliable estimate of σ^2 than is required. Therefore, let us investigate how to choose n_2 so that the relative effort allotted to the estimation of σ_a^2 and σ^2 satisfies a predetermined measure of the relative importance of these components.

If

$$R = \frac{\text{Var } [\hat{\sigma}_a^2]}{\text{Var } [\hat{\sigma}^2]}, \tag{3}$$

where R will be called the relative precision of estimation, then

$$R = \frac{n_1(n_2 - 1)(1 + n_2k)^2 + (n_1 - 1)}{n_2^2(n_1 - 1)}. \tag{4}$$

The expression (4) reduces to a cubic equation in n_2 , which can be solved to indicate the correct choice of n_2 for a balanced allocation, or bounds for the correct solution can be derived in the following manner:

(a) Approximate R by

$$R' = \frac{(n_1 - 1)(n_2 - 1)(1 + n_2k)^2 + (n_1 - 1)}{n_2^2(n_1 - 1)}. \tag{5}$$

Subject to $n_2 > 0$, (5) reduces to the quadratic equation in n_2

$$n_2^2(k^2) - n_2[R' + k(k - 2)] - (2k - 1) = 0 \tag{6}$$

with roots

$$n_2(R') = \frac{[R' + k(k - 2)] \pm \sqrt{[R' + k(k - 2)]^2 - 4k^2(1 - 2k)}}{2k^2}. \tag{7}$$

The larger root will be considered because only very infrequently does the smaller root supply an applicable solution. Using this root, the relative precision actually achieved, R_1, is

$$R_1 = \frac{n_1}{n_1 - 1} R' - \frac{1}{n_2^2(n_1 - 1)} > R'.$$

(b) Approximate R by

$$R'' = \frac{n_1(n_2 - 1)(1 + n_2k)^2 + n_1}{n_2^2(n_1 - 1)}. \tag{8}$$

Subject to $n_1 > 1$, $n_2 > 0$, (8) reduces to the quadratic equation

$$n_2^2(n_1k^2) - n_2[(n_1 - 1)R'' + n_1k(k - 2)] - n_1(2k - 1) = 0 \tag{9}$$

with roots

$$n_2(R'') = \frac{[(n_1 - 1)R'' + n_1k(k - 2)]}{2n_1k^2} \pm \frac{\sqrt{[(n_1 - 1)R'' + n_1k(k - 2)]^2 - 4n_1^2k^2(1 - 2k)}}{2n_1k^2}. \tag{10}$$

Again only the larger root will be considered. The precision achieved for this allocation, R_2, is

$$R_2 = R'' - \frac{1}{n_2^2(n_1 - 1)} < R''.$$

R'' is the more desirable of the approximations since it utilizes the value n_1, and it provides a solution only slightly smaller than the desired precision. For large values of n_1, however, it is readily seen from an examination of the interval about the specified ratio,

$$R_2 = R - \frac{1}{n_2^2(n_1 - 1)} < R < R_1 = \frac{n_1}{n_1 - 1} R - \frac{1}{n_2^2(n_1 - 1)}, \tag{11}$$

that the result from either approximation provides a relative precision very close to R. Thus, for large predetermined n_1, the solution for n_2 is dependent only on the relative sizes of the variance components to be estimated and the relative precision desired for these estimates. In fact, $n_2(R')$ and $n_2(R'')$ are strictly increasing in the range R', $R'' \geq n_1k(2 - k)/(n_1 - 1)$, which exists for the majority of the situations encountered. Increasing n_2 above $n_2(R'')$ only will increase R and place more emphasis on the estimation of σ^2 than is desired, particularly since σ_a^2 usually is the more interesting parameter.

Most importantly, $n_2(R') > n_2(R'')$, so that an immediate check is available on whether or not a balanced allocation exists for a desired relative precision. If $n_2(R') < 2$, then a larger relative precision must be accepted if a balanced allocation is to be used. From (7) it is seen that a real-valued solution of $n_2(R')$ always exists if $k \geq \frac{1}{2}$. For this case, $n_2(R') < 2$ is certain to occur if $R < k(k + 1) + \frac{1}{2}$. When $k < \frac{1}{2}$ and the solution is real, i.e. $R > k[(2 - k) + 2\sqrt{1 - 2k}]$, no balanced allocation exists if simultaneously $R < k(k + 1) + \frac{1}{2}$, a balanced allocation is certain to exist if $R > k(3k + 2)$, and between these bounds the existence is questionable. Obviously for large values of k, balanced allocations exist only when there is a willingness to place much more emphasis on the estimation of σ^2 than σ_a^2 .

In the case of predetermined n_1 , $n_2 = 2$ probably will be accepted when the balanced solution doesn't exist since large R indicates only that a better estimate of σ^2 will be obtained than that which was thought necessary.

(ii) The more important allocation problem occurs when N is predetermined, and n_1 and n_2 must be chosen, subject to $n_1 n_2 = N$, so that the desired relative precision R is achieved.

In (4) the number of primary units n_1 can be replaced by N/n_2 , and the correct allocation of secondary units satisfies the quadratic equation

$$n_2^2(R + Nk^2) - n_2 N[R + k(k - 2)] + [N(1 - 2k) - 1] = 0. \qquad (12)$$

The desired solutions are obtained from

$$n_2(R) = \frac{N[R + k(k - 2)]}{2(R + Nk^2)} \pm \frac{\sqrt{N^2[R + k(k - 2)]^2 - 4[N(1 - 2k) - 1](R + Nk^2)}}{2(R + Nk^2)}, \qquad (13)$$

$$n_1(R) = N \frac{N[R + k(k - 2)]}{2[N(1 - 2k) - 1]} \pm \frac{\sqrt{N^2[R + k(k - 2)]^2 - 4[N(1 - 2k) - 1](R + Nk^2)}}{2[N(1 - 2k) - 1]}. \qquad (14)$$

With this solution, it is easy to check whether a balanced allocation exists for the k at hand and the desired R. Consider first the case where $R + k^2 - 2k$ is positive. If $k > (N - 1)/2N$, $n_2(R)$ is obtained from the solution with the positive radical element, $n_1(R)$ from the solution with the negative radical element. If $k < (N - 1)/2N$, two possibly acceptable allocations exist by the choice of the positive and negative radical elements. If $k = (N - 1)/2N$, the solution, taken directly from (12), is $n_2(R) = [4N^2R - (N - 1)(3N + 1)]/[4NR + (N - 1)^2]$ and $n_1(R) = N/n_2(R)$. If $R + k^2 - 2k$ is negative, the result for $k > (N - 1)/2N$ holds, but there is no allocation for $k \leq (N - 1)/2N$.

If $n_1(R)$ or $n_2(R)$ is less than two, intuitively, it seems clear that there will be some staggered allocations which will provide more desirable estimates of σ_a^2 and σ^2 than the minimum balanced allocation achieved by setting these values at two.

The same development can be extended to balanced, rows and columns, cross-classified designs. Let r be the number of rows, c be the number of columns, σ_r^2 be the variance attributable to the row effects, σ_c^2 be the variance attributable to the column effects, and σ_e^2 be the residual variance. Set $k_1 = \sigma_c^2/\sigma_e^2$ and $k_2 = \sigma_r^2/\sigma_e^2$, then

$$\frac{\text{Var}\,[\hat{\sigma}_r^2]}{\text{Var}\,[\hat{\sigma}_c^2]} = \frac{r^2}{c^2}\left[\frac{(c - 1)(1 + ck_2)^2 + 1}{(r - 1)(1 + rk_1)^2 + 1}\right]. \tag{15}$$

The proper allocation is obtained from

$$r(R) = \frac{N[k_1(k_1 - 2)R - k_2(k_2 - 2)]}{2[Nk_1^2R - (1 - 2k_2)]} \pm \frac{\sqrt{N^2[k_1(k_1 - 2)R - k_2(k_2 - 2)]^2 - 4N[(1 - 2k_1)R - Nk_2^2][Nk_1^2R - (1 - 2k_2)]}}{2[Nk_1^2R - (1 - 2k_2)]}, \tag{16}$$

$$c(R) = \frac{N[k_1(k_1 - 2)R - k_2(k_2 - 2)]}{2[(1 - 2k_1)R - Nk_2^2]} \pm \frac{\sqrt{N^2[k_1(k_1 - 2)R - k_2(k_2 - 2)]^2 - 4N[(1 - 2k_1)R - Nk_2^2][Nk_1^2R - (1 - 2k_2)]}}{2[(1 - 2k_1)R - Nk_2^2]}. \tag{17}$$

The allocation will depend on

$$[(1 - 2k_1)R - Nk_2^2] \quad \text{and} \quad [Nk_1^2R - (1 - 2k_2)], \tag{18}$$

because the signs of the radical must be different in the two solutions. Three sets of conditions exist which affect the choice of r and c. These conditions are functions only of the physical limitations of the estimation problem (N, k_1 , and k_2) and the desired relative precision of estimation (R). Let $k_1(k_1 - 2)R - k_2(k_2 - 2)$ be positive;

(a) If both elements of (18) are positive, then there are two solutions from (16) and (17) which can lead to acceptable allocations.

(b) If the elements of (18) have opposing signs, then there exists one solution from (16) and (17) for an acceptable allocation.

(c) If both elements of (18) are negative, there is no acceptable allocation.

When $k_1(k_1 - 2)R - k_2(k_2 - 2)$ is negative, the results of (a) and (c) are interchanged. In each case, there is the hazard that the solutions for a desired R will be complex values.

In the examination for possible allocations of the nested, fixed-total design and the cross-classified design, it is easier to check for the existence of an allocation by substitutions into (13), (14), (16), and (17) of the parameters of a specific problem than to give general bounds on R. A rough indication of the existence of an allocation can be obtained by replacing k, k_1, and k_2 by estimated values. If data from a pilot experiment of the nested type is available, the estimate

$$\hat{k} = \frac{N' - n_1' - 2}{N' - n_1'}\left[\frac{\hat{\sigma}_a'^2}{\hat{\sigma}'^2} - \frac{2}{n_2'(N' - n_1' - 2)}\right], \tag{19}$$

where the primes indicate values from the pilot data, is unbiased. The estimates of k_1 and k_2 have the same form as (19) where $\hat{\sigma}_a'^2$ and $\hat{\sigma}'^2$ are replaced by $\hat{\sigma}_r'^2$ or $\hat{\sigma}_c'^2$ and $\hat{\sigma}_e'^2$, and $(r' - 1)$ or $(c' - 1)$ replaces $(n_1' - 1)$, $(r' - 1)(c' - 1)$ replaces $N' - n_1'$, and $1/c'$ or $1/r'$ replaces $1/n_2'$. It is the experience of the author that this check indicates a balanced allocation only when the ratio of the estimate of k to the desired value of R is small.

The existence check also provides an estimate of the desired allocation if one is indicated. In these estimated allocations, it is unlikely that any of the solutions using $\hat{k}$ or the true k will be an integer. No investigation has been made as to what should be done in this case, but it seems reasonable that if $n_2(R')$ and $n_2(R'')$ bound an integer, use that integer. If they do not, take the next largest integer; at worst, for known k, this will give an estimate of σ^2 which is better than desired. For $n_2(R)$, $r(R)$, and $c(R)$, take that one of the integers bounding the solution which, when substituted into (6), (16) or (17) most closely approaches the specified R. Like the existence check, the estimated allocations should be regarded only as crude approximations to the desired allocations.

AN EXTENSION FOR THE JOINT ESTIMATION OF p SETS OF VARIANCES IN NESTED ALLOCATIONS

Frequently, as in genetic work, more than one variate is measured on each individual experimental unit. For p such variates it is desired to obtain p sets of estimates of the variances $(\sigma_i^2, \sigma_{ai}^2)$; $i = 1, \cdots, p$.

Since it is often economically unwise to allocate by the variate rather than the whole experimental unit, the allocation must be in terms of a single measure of relative precision for all p sets of estimates.

It may be that there is one particularly important variate $x_{i'}$ which must provide estimates with a relative precision $R_{i'}$. Then the allocation would be decided by using $R_{i'}$ and $k_{i'}$ in the formulae of the preceding section. If, however, one is satisfied with an average relative precision over all variates measured, let

$$\bar{R} = \sum_{i=1}^{p} R_i/p. \tag{20}$$

Then,

$$\bar{R} = \frac{n_1(n_2 - 1)\left[1 + 2n_2\left(\sum_{i=1}^{p} \frac{k_i}{p}\right) + n_2^2\left(\sum_{i=1}^{p} \frac{k_i^2}{p}\right)\right] + (n_1 - 1)}{n_2^2(n_1 - 1)}.$$

Define

$$(k) = \sum_{i=1}^{p} k_i/p, \qquad (k^2) = \sum_{i=1}^{p} k_i^2/p.$$

The two approximations for case (i) with fixed n_1 are

$$n_2(\bar{R}') = \frac{[\bar{R}' + (k^2) - 2(k)] + \sqrt{[\bar{R}' + (k^2) - 2(k)]^2 - 4(k^2)(1 - 2(k))}}{2(k^2)}. \tag{21}$$

$$\begin{aligned} n_2(\bar{R}'') = {} & \frac{[(n_1 - 1)\bar{R}'' + n_1(k^2) - 2n_1(k)]}{2n_1(k^2)} \\ & + \frac{\sqrt{[(n_1 - 1)\bar{R}'' + n_1(k^2) - 2n_1(k)]^2 - 4n_1(k^2)(1 - 2(k))}}{2n_1(k^2)}, \end{aligned} \tag{22}$$

and for case (ii),

$$\begin{aligned} n_2(\bar{R}) = {} & \frac{N[R + (k^2) - 2(k)]}{2[R + N(k^2)]} \\ & + \frac{\sqrt{N^2[R + (k^2) - 2(k)]^2 - 4[N(1 - 2(k)) - 1][R + N(k^2)]}}{2[R + N(k^2)]}. \end{aligned} \tag{23}$$

In general the features of the single variate case apply here, except it should be noted that $(k^2) \geq (k)^2$.

There are two drawbacks to using $\bar{R}$ as a measure of relative precision. If any k_i is exceedingly large, it precludes the existence of a balanced allocation, even though singly each of the remaining $p - 1$ variance sets could be estimated using balanced allocations. The average R does not incorporate any information about the relative importance of the different sets of estimates. A variate of small interest may provide estimates with a low relative precision, while a very important variate may provide estimates with a very large relative precision.

This generalization to p sets of variances obviously does not hold for the cross-classified designs.

AN EXAMPLE

In Table 1, an analysis of variance of plant data from a genetic experiment with inbred lines is presented. The among lines variance

TABLE 1

Analysis of Variance of Four Inbred Lines

Source of variation	D.f.	M.S.
Among lines	3	19.03
Within lines	36	1.02

σ_a^2 is the genetic variance among the plants being studied, and the within line variance, σ^2, is variation caused by the environment. These two variance components determine the coefficient of heritability of the character being studied; thus both must be estimated.

These data can be used to plan two types of experiment for estimating σ_a^2 and σ^2. In the first experiment, it will be assumed that 200 inbred lines are available for testing and ample land is available for setting out any reasonably sized experiment. For the second experiment, it will be assumed that there is only enough land for a 500 plot experiment, but that any number of inbred lines are available for planting.

To check for the existence of balanced allocations, k is estimated using formula (19) with $N' = 40$, $n_1' = 4$, $n_2' = 10$, $\hat{\sigma}_a'^2 = (19.03 - 1.02)/10 = 1.80$, and $\hat{\sigma}'^2 = 1.02$. The estimate, $\hat{k}$, is 1.66. Since $\hat{k}$ is greater than .5, it should be suspected that R will have to be considerably larger than one to obtain a balanced allocation. Let any R value between one and five be acceptable.

(i) Experiment 1: A balanced allocation is indicated if $R \geqq \hat{k}(\hat{k} + 1) + .5 = 4.92$. The upper acceptable limit of R does provide an estimate of a balanced allocation. Substituting $R = 5$ and $\hat{k} = 1.66$ into (7) or (10) gives an estimated allocation of $n_1 = 200$, $n_2 = 2$; an experiment utilizing 400 plots.

(ii) Experiment 2: For this experiment $R + \hat{k}^2 - 2\hat{k} = 4.44$ when $R = 5$, and $\hat{k} > (N - 1)/2N$; the larger root of (13) is used. To the nearest integer values, the estimated allocation is $n_1 = 250$, $n_2 = 2$.

In either experiment, to achieve a more even distribution of the variances of the estimates, an unbalanced allocation would have to be used.

REFERENCES

Cochran, W. G. [1953]. *Sampling Techniques.* John Wiley and Sons, Inc., New York.

Crump, P. P. [1954]. *Optimal Designs to Estimate the Parameters of a Variable Compound Model.* Unpublished Ph.D. Thesis, North Carolina State College, Raleigh, North Carolina.

Gaylor, D. W. [1960]. *The Construction and Evaluation of Some Designs for the Estimation of Parameters in Random Models.* Unpublished Ph.D. Thesis, North Carolina State College, Raleigh, North Carolina.

AUGMENTED DESIGNS WITH ONE-WAY ELIMINATION OF HETEROGENEITY[1]

WALTER T. FEDERER

Cornell University, Ithaca, New York, U.S.A.

INTRODUCTION

One of the principal problems in plant breeding and in biochemical research of new pesticides, herbicides, soil fumigants, drugs, etc., is the evaluation of the new strain or chemical. Efficient experimental designs and efficient screening procedures are necessary in order to make the most efficient use of available resources. In some instances sufficient material of a new strain or a new chemical is available for only one or two observations (plots). Hence, the experimenter should use an experimental design and a screening procedure suitable for these conditions. In other cases, the experimenter may wish to limit his observations to a single observation on the new material. In still other cases (e.g., in physics), a single observation on new material may be desirable because of relatively low variability in the experimental material. Furthermore, it may be desired to combine screening experiments on new material and preliminary testing experiments on promising material. The experimental design should be selected to meet the requirements of such experiments rather than selecting the material and experiments to meet the requirements of the experimental design. The experimental designs described in the present paper were developed to satisfy requirements such as those described above.

The class of experimental designs known as *augmented designs* was introduced by the author in 1955 to fill a need arising in screening new strains of sugar cane and soil fumigants used in growing pineapples[2] (Federer [1956a, 1956b, 1956c, 1958]). An augmented experimental design is any standard design augmented with additional treatments in the complete block, the incomplete block, the row, the column, etc.

[1] Paper No. BU-39 of the Biometrics Unit and Paper No. 341 of the Department of Plant Breeding.

[2] G. N. Rao, Sugar Cane Research Station, Anakapalle, Andra Pradesh, India, has used some of these designs for field lay-outs of sugar cane (personal communication, 7/19/57). J. G. Darroch, Experiment Station, Hawaiian Sugar Planters' Association, Honolulu, Hawaii, expected that eight experiments designed as augmented randomized complete block designs would be harvested in 1957 and several more had been installed (personal communication, 8/23/57).

The construction and randomization procedures will become apparent after consideration of a few specific examples. Analyses for some of these designs have appeared in the literature (Federer [1956a, 1956c]). The purpose of this paper is to present the general approach for all augmented designs with one-way elimination of heterogeneity and to present examples of two specific augmented designs, the augmented randomized complete block design and the augmented balanced lattice design. The general approach for designs with two- and higher-way elimination of heterogeneity will be presented in a forthcoming paper.

Examples of specific designs with additional treatments, unequal replication on treatments, or unequal block sizes have appeared in statistical literature (e.g., Basson, [1959]; Corsten, [1959]; Das, [1958]; Graybill and Pruitt, [1958]; Justensen and Keuls, [1958]; Kishen, [1941]; McIntyre, [1958]; Pearce, [1948]; Yates, [1936b]; Youden and Connor, [1953]). However, a general approach covering the class of augmented designs as well as the others has not been presented. This is done in the present paper.

The randomized complete block design and the one-restrictional lattice designs are well known examples of experimental designs with one-way elimination of heterogeneity. General methods of analyses have been developed for incomplete block designs and for non-orthogonal situations (e.g., Bose and Nair, [1939]; Federer, [1957]; Kempthorne, [1952]; Nair, [1941]; Rao, [1947]; Yates, [1934], [1936a], [1936b], [1938]). Analyses for augmented designs with one-way elimination of heterogeneity are developed along similar lines.

CONSTRUCTION AND RANDOMIZATION

The construction of augmented designs with one-way elimination of heterogeneity is illustrated with examples. Consider first the augmented randomized complete block design. Here there are $N_j = n_{.j}$ plots (experimental units) in each of the $j = 1, \cdots, r$ blocks, with the N_j not necessarily equal; there are two kinds of treatments, treatments repeated r times and occurring once in every block and treatments repeated less than r times (the treatments could appear more than r times in the experiment and the analyses still hold) and hence occurring in only a portion of the blocks. For a large number of situations a number v_r of treatments will occur once in each of the r blocks and a number v_1 of treatments will occur once in one of the r blocks; for convenience call the former group *standards* or *standard treatments* and the latter *new treatments*. The schematic treatment layout for $r = 5$ blocks, $v_r = 4$ standards (A, B, C, D), and $v_1 = 13$ new treatments $(e, f, g, h, i, j, k, l, m, n, o, p, q)$ is:

	Block number				
	1	2	3	4	5
	A	A	A	A	A
	B	B	B	B	B
	C	C	C	C	C
	D	D	D	D	D
	e	h	k	n	p
	f	i	l	o	q
	g	j	m	—	—
N_j	7	7	7	6	6

where the N_j (=6 or 7) were made as nearly equal as possible although this is not necessarily the best grouping for all experimental situations.

The randomization procedure for the augmented randomized complete block design is:

(i) Randomly allot the v_r standards to v_r of the $N_j = n_{.j}$ plots in each block.

(ii) Randomly allot the v_1 new treatments to the remaining plots, which is equivalent to randomly assigning the lower case letters to the new treatments and assigning the letters in order to the remaining plots.

(iii) If a new treatment appears more than once, assign the different entries of the treatment to a complete block at random with the proviso that no treatment occurs more than once in a complete block until that treatment occurs once in each of the complete blocks.

The augmented triple lattice design with $v_{3q} = v_r = 9$ standard treatments (capital letters) and $v_1 = 15$ new treatments (lower case letters) is used to illustrate the construction of augmented incomplete block designs. The schematic lay-out for the treatments is:

	Replicate and incomplete block number								
	1			2			3		
	1	2	3	1	2	3	1	2	3
	A	D	G	A	B	C	A	C	B
	B	E	H	D	E	F	E	D	F
	C	F	I	G	H	I	I	H	G
	j	l	n	p	r	t	u	v	w
	k	m	o	q	s	—	—	—	x
$n_{.jh}$	5	5	5	5	5	4	4	4	5

The same procedure may be used to obtain other incomplete block designs. For example, a fourth group of treatments which would make the above an augmented balanced lattice design, is (for $v_1 = 18$):

	Replicate 4 Incomplete block no.		
	1	2	3
	A F H y	B D I z	C E G α
$n_{.jh}$	4	4	4

The randomization procedure for any augmented incomplete block design for v_r standards, repeated r times, and $v - v_r$ new treatments in b incomplete blocks of size $n_{.jh}$ is:

(i) Randomly allot the groups to the incomplete blocks within each replicate.

(ii) Randomly allot the standards within each group to the $n_{.jh}$ plots of block jh.

(iii) Randomly assign the new treatments to the remaining plots.

(iv) New treatments appearing more than once in the experiment should be randomly allotted to the incomplete blocks with the provisos that the treatment should not appear twice in any one replicate until it has appeared once in all replicates and the treatment should not appear twice in one incomplete block until it has occurred once in all incomplete blocks.

Other incomplete block designs (Federer, [1955], chapters XI and XIII; Kempthorne, [1952], chapters 22, 25, 26) may be used to set up additional augmented incomplete block designs. The procedure is as described above for augmented triple and balanced lattice designs.

ANALYSES

The generalized form of the analysis of variance for an augmented design with one-way elimination of heterogeneity is presented in Table 1; the algebraic developments and the notational definitions for this table are presented in the section entitled Appendix. (This development logically appears in this section, but was relegated to an appendix because of algebraic complexities.) The intrablock treatment means and variances and the interblock treatment means and variances are presented in matrix form. Likewise, the expected values of the mean squares E_e . E_b and E_b' from Table 1 are presented in general form. Utilizing these results, the analysis of variance, adjusted treatment means, and variances for mean differences are presented for the augmented randomized complete block design and for the resolvable augmented balanced lattice design in the following two sections. Analyses for other incomplete block designs may be obtained from the general results given in the Appendix.

TABLE 1

Analysis of Variance for Experimental Designs with One-Way Elimination of Heterogeneity.

Source of variation	df	Sum of squares	Mean squares
Total (uncorrected)	$n_{...}$	$\sum \sum \sum n_{ijh} Y_{ijh}^2$	—
Correction for mean = CF	1	$Y_{...}^2/n_{...}$	—
Among incomplete blocks (ignoring treatments and complete blocks)	$b - 1$	$\sum_j \sum_h \frac{Y_{.jh}^2}{n_{.jh}} - \frac{Y_{...}^2}{n_{...}}$	—
Among treatments (eliminating complete and incomplete block effects)	$v - 1$	$\sum_i \hat{\tau}_i Q_{i..}$	T'
Intrablock error	$n_{...} - b - v + 1$	subtraction	E_e
Complete blocks (eliminating treatments, ignoring incomplete blocks)	$r - 1$	$\sum_j \rho_j^* Q_{.j.}$	—
Incomplete blocks (eliminating treatments and complete blocks)	$b - r$	$\sum_j \sum_h \widehat{(\rho_j + \beta_{jh})} Q_{.jh}$ $-\sum_j \rho_j^* Q_{.j.}$	E_b
Among complete and incomplete blocks (eliminating treatments)	$b - 1$	$\sum_j \sum_h \widehat{(\rho_j + \beta_{jh})} Q_{.jh}$	E_b'

AUGMENTED RANDOMIZED COMPLETE BLOCK DESIGN

For the augmented randomized complete block design, it is simpler algebraically to look at the normal equations for the effects from the intrablock analysis rather than to consider the v formulae given by (3).

The discussion here refers to new treatments appearing only once in the experiment. If the new treatments appear more than once, it may be simpler to obtain solutions from equation (4). The normal equations are:

$$\hat{\mu}: n_{..}\hat{\mu} + \sum_{i=1}^{r} n_{i.}\hat{\tau}_i + \sum_{j=1}^{r} n_{.j}\hat{\rho}_j = Y_{...} ,$$

$$\hat{\rho}_j \ : n_{.j}(\hat{\mu} + \hat{\rho}_j) + \sum_{i=1}^{v} n_{ij}\hat{\tau}_i = Y_{..j} .$$

Standard treatments (in r blocks):

$$r(\hat{\mu} + \hat{\tau}_{ri}) + \sum_{j=1}^{r} \hat{\rho}_j = Y_{ri.} .$$

New treatments (in one of the r blocks):

$$\hat{\mu} + \hat{\rho}_j + \hat{\tau}_{1ij} = Y_{1ij} ,$$

where[3]

$$\sum_{i=1}^{v} \hat{\tau}_i = \sum_{i=1}^{v_r} \hat{\tau}_{ri} + \sum_{i=v_r+1}^{v_r+v_1} \hat{\tau}_{1ij} = \sum \hat{\rho}_j = 0.$$

Solution of these equations for effects results in the following (Federer, [1956a] and [1956c]):

$$\hat{\mu} = \frac{1}{v_r + v_1}\left\{Y_{...} - (r-1)\sum_{i=1}^{v_r} \bar{y}_{ri.} - \sum n_{1j}\hat{\rho}_j\right\},$$

$$\hat{\mu} + \hat{\tau}_{ri} = Y_{ri.}/r = \bar{y}_{ri.} ,$$

$$\hat{\rho}_j = \frac{1}{v_r}\left\{Y_{..j} - \sum_{i=1}^{v_r} \bar{y}_{ri.} - \sum_{k=1}^{n_{1j}} Y_{1kj}\right\} = \frac{1}{v_r}\left\{Y_{r.j} - \sum_{i=1}^{v_r} \bar{y}_{ri.}\right\},$$

($N_j - v_r = n_{1j} =$ number of new treatments in block j)

$$\hat{\mu} + \hat{\tau}_{1ij} = Y_{1ij} - \hat{\rho}_j = Y'_{1ij} .$$

Thus, the new treatment means need to be adjusted for the block in which they appear.

The various estimated variances between treatment means are:

Two standards:

$$V(\bar{y}_{ri.} - \bar{y}_{rk.}) = 2E_e/r \qquad (i \neq k).$$

Two new treatments in the same block:

$$V(Y'_{1ji} - Y'_{1jk}) = 2E_e \qquad (i \neq k).$$

Two new treatments in different blocks:

[3] A somewhat simpler solution could have been obtained by taking $\sum_{i=1}^{v_r} \hat{\tau}_{ri} = 0$ rather than $\sum_{i=1}^{v} \hat{\tau}_i = 0$.

$$V(Y'_{1ij} - Y'_{1hk}) = 2E_e(1 + 1/v_r) \qquad (i \neq k;\, j \neq h).$$

A standard and a new treatment:

$$V(\bar{y}_{ri.} - Y'_{1jk}) = E_e\left(1 + \frac{1}{r} + \frac{1}{v_r} - \frac{1}{bv_r}\right) \qquad (i \neq k),$$

where E_e is the error mean square obtained from an analysis of variance on the standards alone.

The following numerical example was constructed for ease of computation ($r = 3$, $v_r = 3$, $v_1 = 2$):

	Replicate number			
	1	2	3	
	$Y_{311} = 9$	$Y_{312} = 6$	$Y_{313} = 12$	
	$Y_{321} = 5$	$Y_{322} = 6$	$Y_{323} = 10$	
	$Y_{331} = 7$	$Y_{332} = 6$	$Y_{333} = 11$	
	$Y_{141} = 13$	$Y_{152} = 10$	——	
Rep. Total	$Y_{..1} = 34$	$Y_{..2} = 28$	$Y_{..3} = Y_{3.3} = 33$	$Y_{...} = 95$
Total for Standards	$Y_{3.1} = 21$	$Y_{3.2} = 18$		$Y_{3..} = 72$
Standard Totals	$Y_{31.} = 27$	$Y_{32.} = 21$	$Y_{33.} = 24$	

Applying the formulae obtained above we find that $\hat{\mu} = 10$, $\hat{\tau}_1 = \hat{\tau}_{31} = -1$, $\hat{\tau}_2 = \hat{\tau}_{32} = -3$, $\hat{\tau}_3 = \hat{\tau}_{33} = -2$, $\hat{\tau}_4 = \hat{\tau}_{11} = 4$, $\hat{\tau}_5 = \hat{\tau}_{12} = 2$, $\hat{\rho}_1 = -1$, $\hat{\rho}_2 = -2$, $\hat{\rho}_3 = 3$.

In the analysis of variance table we obtain:[4]

Source of variation	df	ss	ms
Total (corrected for mean)	10	76.5454	—
Blocks (ignoring treatment)	2	27.5454	
Treatments (eliminating blocks)	4	45	11.25
Error (elim. tr. and bl.)	4	4	1
Block (eliminating treatment)	2	42	21
Treatment (ignoring blocks)	4	30.5454	—

The error sum of squares equals 4 as it should since $\epsilon_{311} = +1$, $\epsilon_{321} = -1$, $\epsilon_{312} = -1$, and $\epsilon_{322} = +1$ was used in constructing the

[4] The error (eliminating treatment and block) and the block (eliminating treatment) sums of squares may be obtained from the analysis of variance on the yields for the standards.

example and $1^2 + (-1)^2 + (-1)^2 + 1^2 = 4$. Also, from the formulae for variances,

$$V(\bar{y}_{3i.} - \bar{y}_{3k.}) = 2(1)/3 = 2/3,$$
$$V(Y'_{141} - Y'_{152}) = 2(1)(1 + 1/3) = 8/3,$$
$$V(\bar{y}_{3i.} - Y'_{1kj}) = (1)(1 + 1/3 + 1/3 - 1/9) = 14/9.$$

Likewise, it is possible to partition the treatment (eliminating block) sum of squares with $v_r + v_1 - 1$ degrees of freedom into the following orthogonal contrasts:[5]

Among standards ($v_r - 1$ *d.f.*)

$$\sum_{i=1}^{v_r} \frac{Y_{ri.}^2}{r} - \frac{Y_{r..}^2}{rv_r} = 6.$$

Among new treatments within a block ($\sum n_{ij} - r$ *d.f.*):

$$\sum_{j=1}^{r} \left\{ \sum_{i=1}^{n_{1j}} Y_{1ji}^2 - \left(\sum_{i=1}^{n_{1j}} Y_{1ji} \right)^2 / n_{1j} \right\} = 0.$$

Standards vs. new treatments in block j (r *d.f.*):

$$\sum_{j=1}^{r} \left\{ n_{1j} \sum_{i=1}^{v_r} Y_{rij} - v_r \sum_{i=1}^{n_{1j}} Y_{1ij} \right\}^2 / v_r n_{1j}(v_r + n_{1j})$$
$$= \frac{(21 - 3(13))^2}{3(1)(3 + 1)} + \frac{(18 - 3(10))^2}{3(1)(3 + 1)} + 0 = 39.$$

Thus, $6 + 0 + 39 = 45$ = treatment (eliminating block) sum of squares.

AUGMENTED BALANCED LATTICE DESIGN

The balanced lattice design, or its equivalent, the balanced incomplete block design, is described in various places (e.g., Federer, [1955], sections XI–3.3, XI–4, and XIII–2.1; Kempthorne, [1952], Chapters 22, 23, and 26). By including additional treatments in the incomplete blocks, an augmented balanced lattice design (ABLD) is formed. The simplest ABLD is the one formed by including new treatments only once in the experiment. If some of the new treatments appear in more than one incomplete block, the computations become more complex, but the general results given in the Appendix apply.

For the ABLD, it appears that the restriction $\sum_{i=1}^{v_r} \hat{\tau}_i = 0$ results in a simpler solution than the restriction $\sum_{i=1}^{v_r} \hat{\tau}_i + \sum_{i=v_r+1}^{v} \hat{\tau}_i = \sum_{i=1}^{v} \hat{\tau}_i = 0$ (the first v_r treatments are standards and the remaining v_r treatments are new ones). Using the $\sum_{i=1}^{v_r} \hat{\tau}_i = 0$ as the restriction, it must be remembered that the differences between treatment means are unbiased but that the treatment means themselves are biased considering all v treatments as fixed effects with

[5] Suggested by a referee.

$$\frac{1}{v}\sum_{i=1}^{v}(\mu_{i..}-\mu)=\frac{1}{v}\sum_{i=1}^{v}\tau_i=0,$$

where $\mu_{i..}$ equals true treatment mean and $\mu = \sum \mu_{i..}/v$. In most situations, interest lies in differences between means rather than in the means themselves.

Intrablock analysis:

The intrablock analysis of the ABLD with the standards in a balanced incomplete block design for $v_r = k^2$, $r = k + 1$, and $b = k(k + 1)$ incomplete blocks of size k, is relatively simple computationally for the standards in $k + 1$ replicates and the new treatments included once. Using the v equations obtained from formula (3), add $1/k$ times the sum of the normal equations for the new treatments which appear in an incomplete block with the ith standard, to the normal equation for the ith standard. Performing the same operations on the normal equations for all standards results in a set of equations involving only the effects and the yields associated with the standards. The resulting equation for the ith standard treatment after using the equation $\sum_{i=1}^{k^2}\hat{\tau}_i = 0$, is

$$\left[\frac{r(k-1)+\lambda}{k}\right]\hat{\tau}_i = Y_{i..} - \sum_{j=1}^{k+1}\sum_{h=1}^{k} n_{ijh}\bar{y}_{r.jh} = Q_{i..}\,,$$

where $\bar{y}_{r.jh}$ equals jhth incomplete block mean on standards only and $\lambda = 1 =$ number of times any two standards appear together in an incomplete block in this particular ABLD. With estimated treatment effects for the standards it is possible to solve for the remaining $v - v_r = v_1\ \hat{\tau}_i$ for the new treatments simply by substituting the $\hat{\tau}_i$ for the standards and solving the $n_{.jh} - k$ equations in the jhth block.

Since $\hat{\rho}_j = \bar{y}_{r.j.} - \bar{y}_r$ (where $\bar{y}_{r.j.} =$ jth complete block mean and $\bar{y}_r =$ overall mean obtained on the yields of the standard treatments alone), the $\hat{\beta}_{jh}$ may be obtained from formula (8) reduced as follows:

$$k\hat{\beta}_{fe} - \frac{1}{k+1}\sum_{i=1}^{k^2} n_{ife}\sum_{j=1}^{k+1}\sum_{h=1}^{k} n_{ijh}\hat{\beta}_{jh}$$

$$= Y_{r.fe} - \sum_{i=1}^{k^2} n_{ife}\bar{y}_{i..} - k(\bar{y}_{r.j.} - \bar{y}_r) = Q'_{.fe}\,.$$

With the additional equation $\sum_{h=1}^{k}\hat{\beta}_{jh} = 0$, we find that the solutions are:

$$\hat{\beta}_{jh} = \frac{k+1}{k^2}Q'_{.jh}\,,$$

which is the usual solution for a resolvable balanced lattice design with the parameters $v = k^2$, $k =$ block size, $r = k + 1$, $b = k(k + 1)$ and

$\lambda = 1$. The above solutions are used in the computation of the sums of squares in the analysis of variance.

The treatment mean adjusted for complete block and for incomplete block effects is $\hat{\mu} + \hat{\tau}_.$ and the variance of a difference between two adjusted means in this ABLD is one of the following:

V (difference between intrablock means of two standards
$$= \hat{\mu} + \hat{\tau}_i - \hat{\mu} - \hat{\tau}_g = \hat{\tau}_i - \hat{\tau}_g) = 2\sigma_\epsilon^2/k,$$
V (difference between a standard and new treatment occurring in the same incomplete block)
$$= \sigma_\epsilon^2(1/k + (k^2 + k + 1)/k^2 - 2/k^2) = \sigma_\epsilon^2(k^2 + 2k - 1)/k^2,$$
V (difference between a standard and new treatment not occurring in the same incomplete block)
$$= \sigma_\epsilon^2(1/k + (k^2 + k + 1)/k^2) = \sigma_\epsilon^2[(k + 1)/k^2],$$
V (difference between two new treatments in the same incomplete block)
$$= \sigma_\epsilon^2\{2(k^2 + k + 1)/k^2 - 2(k + 1)/k\} = 2\sigma_\epsilon^2\ ,$$
V (difference between two new treatments in the same replicate but not in the same incomplete block)
$$= 2\sigma_\epsilon^2(k^2 + k + 1)/k^2,$$
V (difference between two new treatments not in the same replicate)
$$= \sigma_\epsilon^2(2(k^2 + k + 1)/k^2 - 2/k^3) = 2\sigma_\epsilon^2(k^3 + k^2 + k - 1)/k^3.$$

The n^{gi} in formulae (4) to (6) may be obtained from the above variances. The intrablock mean square E_e is an estimate of σ_ϵ^2 . Hence, the estimated variances are obtained simply by substituting E_e for σ_ϵ^2 in the above six variances.

The following numerical example was selected for ease of computation. The non-randomized layout is presented below along with the computations:

	Replicate I	Replicate II	Replicate III
	$Y_{111} = 4$ $Y_{312} = 14$	$Y_{121} = 11$ $Y_{222} = 8$	$Y_{131} = 7$ $Y_{232} = 10$
	$Y_{211} = 5$ $Y_{412} = 13$	$Y_{421} = 10$ $Y_{322} = 15$	$Y_{331} = 11$ $Y_{432} = 12$
	$Y_{511} = 7$	$Y_{722} = 9$	
	$Y_{611} = 9$		
Totals	$Y_{.11} = 25$ $Y_{.12} = 27$ $Y_{.1.} = 52$	$Y_{.21} = 21$ $Y_{.22} = 32$ $Y_{.2.} = 53$	$Y_{.31} = 18$ $Y_{.32} = 22$ $Y_{.3.} = 40$
Totals standards	$Y_{3.11} = 9$ $Y_{3.12} = 27$ $Y_{3.1.} = 36$	$Y_{3.21} = 21$ $Y_{3.22} = 23$ $Y_{3.2.} = 44$	$Y_{3.31} = 18$ $Y_{3.32} = 22$ $Y_{3.3.} = 40$

Treatment totals $Y_{1..} = 22$, $Y_{2..} = 23$, $Y_{3..} = 40$, $Y_{4..} = 35$, $Y_{5..} = 7$, $Y_{6..} = 9$, $Y_{7..} = 9$. Grand total $= Y_{...} = 145$; grand total on standards $= Y_{3...} = 120$.

$$Q_{1..} = -45/12, \qquad Q_{2..} = -59/12, \qquad Q_{3..} = 82/12,$$

$$Q_{4..} = 0, \qquad Q_{5..} = 9/12, \qquad Q_{6..} = 33/12,$$

$$Q_{7..} = -20/12, \qquad Q'_{.11} = -4, \qquad Q'_{.12} = 4,$$

$$Q'_{.21} = 0, \qquad Q'_{.22} = 0, \qquad Q'_{.31} = -8/3,$$

$$Q'_{.32} = 8/3.$$

$$\hat{\mu} = 120/12 = 10, \qquad \hat{\rho}_1 = (36/4) - 10 = -1,$$

$$\hat{\rho}_2 = (44/4) - 10 = 1, \qquad \hat{\rho}_3 = (40/4) - 10 = 0.$$

$$\begin{bmatrix} \tau_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \\ \hat{\tau}_4 \\ \hat{\tau}_5 \\ \hat{\tau}_6 \\ \hat{\tau}_7 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 0 & 0 & 1/4 & 1/4 & 0 \\ 0 & 1/2 & 0 & 0 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 1/4 & 1/4 & 0 & 0 & 7/4 & 3/4 & 1/8 \\ 1/4 & 1/4 & 0 & 0 & 3/4 & 7/4 & 1/8 \\ 0 & 1/4 & 1/4 & 0 & 1/8 & 1/8 & 7/4 \end{bmatrix} \begin{bmatrix} -45/12 \\ -59/12 \\ 82/12 \\ 0 \\ 9/12 \\ 33/12 \\ -20/12 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \\ +3 \\ 0 \\ 1 \\ 3 \\ -2 \end{bmatrix}.$$

$$\begin{bmatrix} \hat{\beta}_{11} \\ \hat{\beta}_{12} \\ \hat{\beta}_{21} \\ \hat{\beta}_{22} \\ \hat{\beta}_{31} \\ \hat{\beta}_{32} \end{bmatrix} = \begin{bmatrix} 3/4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3/4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3/4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3/4 \end{bmatrix} \begin{bmatrix} -4 \\ 4 \\ 0 \\ 0 \\ -8/3 \\ 8/3 \end{bmatrix} = \begin{bmatrix} -3 \\ 3 \\ 0 \\ 0 \\ -2 \\ 2 \end{bmatrix}.$$

Using the above, the sums of squares in the analysis of variance table for the standard treatments are computed in the usual manner (see Federer, [1955], page 342) and are:

Source of variation	df	ss	ms
Total	12	1330	—
CFM	1	1200	—
Replicates	2	8	—
Standards (ign. blocks)	3	238/3	—
Blocks (elim. treatment)	3	104/3	$104/9 = E_b$
Intrablock error	3	8	$8/3 = E_e$

The analysis of variance on all treatments is:

Source of variation	df	ss	ms
Total	15	1541	—
CFM	1	4205/3	—
Replicates (ign. tr. and bl.)	2	54/5 = 10.8	—
Blocks (elim. reps; ign. tr.)	3	4447/60	—
Treatments (elim. bl. and reps)	6	557/12	557/72
Intrablock error	3	8	$8/3 = E_e$
Blocks (elim. tr. and reps)	3	104/3	$104/9 = E_b$

The various variances of a mean difference between two adjusted treatment means are:

Variance of difference between two standards

$$2E_e/k = 8/3.$$

Variance of difference between standard and new treatment in same incomplete block

$$E_e\left(\frac{k^2 + 2k - 1}{k^2}\right) = \left(\frac{8}{3}\right)\left(\frac{4 + 4 - 1}{4}\right) = 14/3.$$

Variance of difference between standard and new treatment not in same incomplete block

$$E_e\left(\frac{k + 1}{k}\right)^2 = \frac{8}{3}\left(\frac{3}{2}\right)^2 = 6.$$

Variance of difference between two new treatments in same incomplete block

$$2E_e = 16/3.$$

Variance of difference between two new treatments in same replicate but different incomplete blocks

Not applicable for this example.

Variance of difference between two new treatments in different replicates

$$2E_e\left(\frac{k^3 + k^2 + k - 1}{k^3}\right) = 2\left(\frac{8}{3}\right)\left(\frac{8 + 4 + 2 - 1}{8}\right) = 26/3.$$

Interblock analysis:

The amount of intrablock information is $w = 1/E_e = \frac{3}{8}$. In order to obtain the amount of intrablock information, we first need to obtain the expectation of E_b from formula (13). After substituting in the various values, a rather surprising result is obtained in that E_b in the ABLD with new treatments occurring only once has the same expectation, $\sigma_\epsilon^2 + k^2\sigma_\beta^2 (k + 1)$, as for the standard balanced lattice design. Therefore, $w' = k/[(k + 1)E_b - E_e] = 2/[3(104/9) - 8/3] = 1/16$. With these weights, we now proceed to the computation of the n_{ig} and $Z_{i..}$ from formulae (21) and (23).[6] Thus $\tau_1^* = -1.440$, $\tau_2^* = -2.132$, $\tau_3^* = 3.132$, $\tau_4^* = .440$, $\tau_5^* = .143$, $\tau_6^* = 2.143$, $\tau_7^* = -2.000$. The variances of a difference between any two τ_i^* may be obtained from formula (24). Thus, the variance of

$$\tau_1^* - \tau_2^* = \frac{998}{819} + \frac{1297}{1071} - \left(\frac{-5}{1071}\right) - \left(\frac{-10}{819}\right) = 2.446.$$

SUMMARY

An augmented experimental design is any standard experimental design to which additional treatments (new treatments) have been added to those (standard treatments) appearing in the standard experimental design. The additional treatments require enlargement of the complete block, the incomplete block, row, column, etc. The groupings in an augmented design may be of unequal size. The construction and randomization procedures, and the general method of analysis have been given for all augmented experimental designs with one-way elimination of heterogeneity from the experimental area. The general results are illustrated algebraically and arithmetically with two examples, an augmented randomized complete block design and an augmented balanced lattice design. Analyses with and without recovery of interblock information are considered. Some discussion of unequal incomplete block sizes is given.

[6] The suggestions leading up to equation (26) were not followed in the computations; the above yields approximately the same results for this example. From equation (26) $w' = 9(53)/(53(104)-23(24)) = .096$.

$$\begin{bmatrix}
\frac{7w}{4}+\frac{19w'}{30} & \frac{-w}{4}-\frac{11w'}{30} & \frac{-w}{2}-\frac{7w'}{60} & \frac{-w}{2}-\frac{7w'}{60} & \frac{w'-3w}{12} & \frac{w'-3w}{12} & \frac{-w'}{5} \\
\frac{-w}{4}-\frac{11w'}{30} & \frac{23w}{12}+\frac{7w'}{15} & \frac{-w}{3}-\frac{17w'}{60} & \frac{-w}{2}-\frac{7w'}{60} & \frac{w'-3w}{12} & \frac{w'-3w}{12} & \frac{2w'-5w}{15} \\
\frac{-w}{2}-\frac{7w'}{60} & \frac{-w}{3}-\frac{17w'}{60} & \frac{5w}{3}+\frac{43w'}{60} & \frac{-w}{2}-\frac{7w'}{60} & \frac{-w'}{6} & \frac{-w'}{6} & \frac{2w'-5w}{15} \\
\frac{-w}{2}-\frac{7w'}{60} & \frac{-w}{2}-\frac{7w'}{60} & \frac{-w}{2}-\frac{7w'}{60} & \frac{3w}{2}+\frac{53w'}{60} & \frac{-w'}{6} & \frac{-w'}{6} & \frac{-w'}{5} \\
\frac{w'-3w}{12} & \frac{w'-3w}{12} & \frac{-w'}{6} & \frac{-w'}{6} & \frac{9w+w'}{12} & \frac{w'-3w}{12} & 0 \\
\frac{w'-3w}{12} & \frac{w'-3w}{12} & \frac{-w'}{6} & \frac{-w'}{6} & \frac{w'-3w}{12} & \frac{9w+w'}{12} & 0 \\
\frac{-w'}{5} & \frac{2w'-5w}{15} & \frac{2w'-5w}{15} & \frac{-w'}{5} & 0 & 0 & \frac{10w+2w'}{15}
\end{bmatrix}
\begin{bmatrix}
\tau^*_1 \\ \tau^*_2 \\ \tau^*_3 \\ \tau^*_4 \\ \tau^*_5 \\ \tau^*_6 \\ \tau^*_7
\end{bmatrix}$$

$$
= \left[\begin{array}{llll}
Z_{1..} = wQ_{1..} + w'\{\bar{y}_{.11} + \bar{y}_{.21} + \bar{y}_{.31} - \bar{y}_{.1.} - \bar{y}_{.2.} - \bar{y}_{.3.}\} & = \dfrac{-15}{32} - \dfrac{211}{1960} & = \dfrac{-1561}{960} \\[2ex]
Z_{2..} = wQ_{2..} + w'\{\bar{y}_{.11} + \bar{y}_{.22} + \bar{y}_{.32} - \bar{y}_{.1.} - \bar{y}_{.2.} - \bar{y}_{.3.}\} & = \dfrac{-59}{32} - \dfrac{81}{960} & = \dfrac{-1851}{960} \\[2ex]
Z_{3..} = wQ_{3..} + w'\{\bar{y}_{.12} + \bar{y}_{.22} + \bar{y}_{.31} - \bar{y}_{.1.} - \bar{y}_{.2.} - \bar{y}_{.3.}\} & = \dfrac{82}{32} + \dfrac{234}{960} & = \dfrac{2694}{960} \\[2ex]
Z_{4..} = wQ_{4..} + w'\{\bar{y}_{.12} + \bar{y}_{.21} + \bar{y}_{.32} - \bar{y}_{.1.} - \bar{y}_{.2.} - \bar{y}_{.3.}\} & = 0 + \dfrac{172}{480} & = \dfrac{344}{960} \\[2ex]
Z_{5..} = wQ_{5..} + w'\{\bar{y}_{.11} - \bar{y}_{.1.}\} & = \dfrac{9}{32} - \dfrac{29}{192} & = \dfrac{125}{960} \\[2ex]
Z_{6..} = wQ_{6..} + w'\{\bar{y}_{.11} - \bar{y}_{.1.}\} & = \dfrac{33}{32} - \dfrac{29}{192} & = \dfrac{845}{960} \\[2ex]
Z_{7..} = wQ_{7..} + w'\{\bar{y}_{.22} - \bar{y}_{.2.}\} & = \dfrac{-20}{32} + \dfrac{1}{240} & = \dfrac{-596}{960}
\end{array}\right].
$$

The solutions for the τ_i^* are:

$$\begin{bmatrix} \tau_1^* \\ \tau_2^* \\ \tau_3^* \\ \tau_4^* \\ \tau_5^* \\ \tau_6^* \\ \tau_7^* \end{bmatrix} = \begin{bmatrix} \frac{998}{819} & \frac{-5}{1071} & \frac{5}{663} & 0 & \frac{15133}{27846} & \frac{15133}{27846} & \frac{839}{13923} \\ \frac{-10}{819} & \frac{1297}{1071} & \frac{-5}{663} & 0 & \frac{14923}{27846} & \frac{14923}{27846} & \frac{7559}{13923} \\ \frac{10}{819} & \frac{5}{1071} & \frac{811}{663} & 0 & \frac{2003}{27846} & \frac{2003}{27846} & \frac{7729}{13923} \\ \frac{-50}{408681} & \frac{3965}{534429} & \frac{-1585}{330837} & \frac{608}{499} & \frac{932927}{13895154} & \frac{932927}{13895154} & \frac{417811}{6947577} \\ \frac{10}{21} & \frac{10}{21} & 0 & 0 & \frac{30}{7} & \frac{34}{21} & \frac{5}{21} \\ \frac{10}{21} & \frac{10}{21} & 0 & 0 & \frac{34}{21} & \frac{30}{7} & \frac{5}{21} \\ 0 & \frac{25}{51} & \frac{25}{51} & 0 & \frac{25}{102} & \frac{25}{102} & \frac{220}{51} \end{bmatrix} \cdot \begin{bmatrix} Z_{1..} = \frac{-1561}{960} \\ Z_{2..} = \frac{-1851}{960} \\ Z_{3..} = \frac{2694}{960} \\ Z_{4..} = \frac{344}{960} \\ Z_{5..} = \frac{125}{960} \\ Z_{6..} = \frac{845}{960} \\ Z_{7..} = \frac{-596}{960} \end{bmatrix}$$

APPENDIX

The generalized analysis of variance is developed below for all designs with one-way elimination of heterogeneity. Such designs as the randomized complete block and incomplete block designs which are resolvable (the v treatments occur together in a complete block) and which are non-resolvable (v treatments occur in b incomplete blocks of size k, for $k < v$), are considered.

To be completely general, let the ith one of the v treatments be replicated r_i times in the b incomplete blocks of size $n_{.jh}$. Let the yield of the ijhth observation be expressed by

$$Y_{ijh} = n_{ijh}(\mu + \tau_i + \rho_j + \beta_{jh} + \epsilon_{ijh}), \tag{1}$$

where $i = 1, \cdots, v$ = number of treatments; $j = 1, \cdots, r$ = number of complete blocks; $h = 1, \cdots, k_j$ = number of incomplete blocks in the jth complete block; $n_{ijh} = 1$ if ith treatment occurs in the hth incomplete block of the jth complete block and zero otherwise;[7] $n_{.jh}$ = number of treatments in hth incomplete block of the jth complete block; $n_{.j.} = v_j$ = number of treatments in the jth complete block; $n_{i..} = r_i$ = number of replicates of the ith treatment;

$$n_{...} = \sum_{i=1}^{v} r_i = \sum_{j=1}^{r} v_j = \sum_{i=1}^{v} \sum_{j=1}^{r} \sum_{h=1}^{k_j} n_{ijh}\ ; \quad \mu_{ijh} = \mu + \tau_i + \rho_j + \beta_{jh}\ ;$$

$$\mu = \text{a general mean effect} = \frac{1}{v} \sum_{i=1}^{v} \mu_{i..} = \frac{1}{r} \sum \mu_{.j.}$$

$$= \frac{1}{n_{...}} \sum_{i} \sum_{j} \sum_{h} n_{ijh}\mu_{ijh} = \frac{1}{b = \sum k_j} \sum_{y=1}^{r} \sum_{h=1}^{k_j} \mu_{.jh}$$

where k_j = number of incomplete blocks in the jth complete block; $\tau_i = \mu_{i..} - \mu$ = a treatment effect; $\rho_j = \mu_{.j.} - \mu$ = a complete block effect; $\beta_{jh} = \mu_{.jh} - \mu_{.j.}$ = an incomplete block effect; and ϵ_{ijh} are random independent effects with mean zero and common variance σ_ϵ^2. These definitions imply $\sum_{i=1}^{v} \tau_i = \sum_{j=1}^{r} \rho_j = \sum_{h=1}^{k_j} \beta_{jh} = 0$. Other definitions for the effects are permissible.

Intrablock Analysis

The least squares estimates of effects for the above linear model (intrablock analysis) are obtained by minimizing the residual sum of squares:

$$\sum_{i=1}^{v} \sum_{j=1}^{r} \sum_{h=1}^{k_j} n_{ijh}\epsilon_{ijh}^2 = \sum_{i} \sum_{j} \sum_{h} n_{ijh}(Y_{ijh} - \mu - \tau_i - \rho_j - \beta_{jh})^2. \tag{2}$$

[7] To be completely general, n_{ijh} could be the number of times treatment i occurs in block jh instead of 1 or zero, but this was not done here.

Equating to zero each of the partial derivatives of the above residual sum of squares with respect to μ, τ_i , ρ_j , and β_{jh} results in the following normal equations:

$$\mu : n_{...}\hat{\mu} + \sum_{i=1}^{v} \hat{\tau}_i n_{i..} + \sum_{j=1}^{r} \hat{\rho}_j n_{.j.} + \sum_{j=1}^{r} \sum_{h=1}^{k_j} n_{.jh}\hat{\beta}_{jh} = Y_{...}$$
$$= \text{grand total},$$

$$\tau_i : n_{i..}(\hat{\mu} + \hat{\tau}_i) + \sum_{j=1}^{r} n_{ij.}\hat{\rho}_j + \sum_{j=1}^{r} \sum_{h=1}^{k_j} n_{ijh}\hat{\beta}_{jh} = Y_{i..}$$
$$= i\text{th treatment total},$$

$$\rho_j : n_{.j.}(\hat{\mu} + \hat{\rho}_j) + \sum_{i=1}^{v} n_{ij.}\hat{\tau}_i + \sum_{h=1}^{k_j} n_{.jh}\hat{\beta}_{jh} = Y_{.j.}$$
$$= j\text{th complete block total},$$

$$\beta_{jh} : n_{.jh}(\hat{\mu} + \hat{\rho}_j + \hat{\beta}_{jh}) + \sum_{i} n_{ijh}\hat{\tau}_i = Y_{.jh}$$
$$= jh\text{th incomplete block total}.$$

In the τ_g equation, substitute for $\hat{\mu} + \hat{\rho}_j + \hat{\beta}_{jh}$ from the β_{jh} equations to obtain:

$$n_{g..}\hat{\tau}_g - \sum_{j} \sum_{h} \frac{n_{gjh}}{n_{.jh}} \sum_{i} n_{ijh}\hat{\tau}_i = Y_{g..} - \sum_{j} \sum_{h} n_{gjh}\bar{y}_{.jh} = Q_{g..} \qquad (3)$$

When $n_{g..} = r$ and $n_{.jh} = k$ the above equation becomes:

$$r\hat{\tau}_g - \frac{1}{k} \sum_{i} \hat{\tau}_i \sum_{j} \sum_{h} n_{gjh}n_{ijh} = Q_{g..}$$

where $\sum_j \sum_h n_{gjh}n_{ijh} = \lambda_{gi}$ = number of times the gth treatment occurs with the ith treatment in all incomplete blocks. For balanced lattices $\lambda_{gi} = \lambda$ a constant for $i \neq g$ and the solution for $\hat{\tau}_g$ is (for $\sum \hat{\tau}_i =$ zero):

$$\hat{\tau}_g = kQ_{g..}/(kr - r + \lambda) = kQ_{g..}(v - 1)/rv(k - 1)$$

as given by Federer ([1955], formula XIII–2, where $Q_{.i/k} = Q_{g..}$). For the general case we can add $d_g \sum \hat{\tau}_i = 0$, where $d_g \neq 0$ for at least one value of g, to each of the v equations in the $\hat{\tau}_i$; adding $d_g \sum \hat{\tau}_i$ to each of the equations in (3) results in:

$$n_{g..}\hat{\tau}_g - \sum_{j} \sum_{h} \frac{n_{gjh}}{n_{.jh}} \sum_{i} \hat{\tau}_i(n_{ijh} - d_g) = Q_{g..} \; .$$

We have v equations and v unknowns and the problem is to solve the equations. In matrix notation the $v + 1$ equations from (3) plus the equation $\sum \hat{\tau}_i = 0$ is:

$$\begin{bmatrix} n_{1..} - \sum\sum \frac{n_{1jh}^2}{n_{.jh}} & -\sum\sum \frac{n_{1jh}n_{2jh}}{n_{.jh}} & \cdots & -\sum\sum \frac{n_{1jh}n_{vjh}}{n_{.jh}} & 1 \\ -\sum\sum \frac{n_{2jh}n_{1jh}}{n_{.jh}} & n_{2..} - \sum\sum \frac{n_{2jh}^2}{n_{.jh}} & \cdots & -\sum\sum \frac{n_{2jh}n_{vjh}}{n_{.jh}} & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -\sum\sum \frac{n_{vjh}n_{1jh}}{n_{.jh}} & -\sum\sum \frac{n_{vjh}n_{2jh}}{n_{.jh}} & \cdots & n_{v..} - \sum\sum \frac{n_{vjh}^2}{n_{.jh}} & 1 \\ 1 & 1 & & 1 & 0 \end{bmatrix}$$

$$\cdot \begin{bmatrix} \hat{\tau}_1 \\ \hat{\tau}_2 \\ \cdots \\ \hat{\tau}_v \\ 0 \end{bmatrix} = \begin{bmatrix} Q_{1..} = Y_{1..} - \sum\sum n_{1jh}\bar{y}_{.jh} \\ Q_{2..} = Y_{2..} - \sum\sum n_{2jh}\bar{y}_{.jh} \\ \cdots \qquad\qquad \cdots \\ Q_{v..} = Y_{v..} - \sum\sum n_{vjh}\bar{y}_{.jh} \\ 0 \end{bmatrix}.$$

Again in matrix notation the solution for the $\hat{\tau}_i$ are obtained as:

$$\begin{bmatrix} \hat{\tau}_1 \\ \hat{\tau}_2 \\ \cdots \\ \hat{\tau}_v \end{bmatrix} = \begin{bmatrix} n^{11} & n^{12} & \cdots & n^{1v} \\ n^{21} & n^{22} & \cdots & n^{2v} \\ \cdots & \cdots & \cdots & \cdots \\ n^{v1} & n^{v2} & \cdots & n^{vv} \end{bmatrix} \begin{bmatrix} Q_{1..} \\ Q_{2..} \\ \cdots \\ Q_{v..} \end{bmatrix}$$

where n^{ig} are the elements of the inverse matrix. The solution for $\hat{\tau}_g$ is

$$\hat{\tau}_g = \sum_{i=1}^{v} n^{gi} Q_{i..} \, , \tag{4}$$

the variance of $\hat{\tau}_g$ is

$$V(\hat{\tau}_g) = n^{gg}\sigma_\epsilon^2 \, , \tag{5}$$

and the variance of a difference between $\hat{\tau}_{g'}$ and $\hat{\tau}_g$ is (Nair, [1941]):

$$V(\hat{\tau}_{g'} - \hat{\tau}_g) = \sigma_\epsilon^2 (n^{g'g'} + n^{gg} - n^{g'g} - n^{gg'}). \tag{6}$$

In the ρ_j equations, substitute for $\hat{\mu} + \hat{\tau}_i$ from $\hat{\tau}_i$ equation, thus for $j = f$:

$$n_{.f.}\hat{\rho}_f - \sum_i \frac{n_{if.}}{n_{i.}} \sum_j \hat{\rho}_j n_{ij.} - \sum_i \frac{n_{if.}}{n_{i..}} \sum_j \sum_h n_{ijh}\hat{\beta}_{jh} + \sum_h n_{.fh}\hat{\beta}_{fh}$$

$$= Y_{.f.} - \sum_i n_{if.}\bar{y}_{i..} = Q_{.f.} \, . \tag{7}$$

The solutions for the ρ_j and β_{jh} must be obtained jointly since they are not orthogonal and since three sets of unknowns are present. Equations involving the τ_i only are possible here because the incomplete blocks in the complete blocks can be considered as b incomplete blocks. For the solution we proceed to the β_{jh} equations and substitute for $\hat{\mu} + \hat{\tau}_i$ to obtain (for $j = f$ and $h = e$):

$$n_{.fe}(\hat{\rho}_f + \hat{\beta}_{.fe}) - \sum_i \frac{n_{ife}}{n_{i..}} \sum_j \sum_h n_{ijh}(\hat{\rho}_j + \hat{\beta}_{jh}) \\ = Y_{.fe} - \sum_i n_{ife}\bar{y}_{i..} = Q_{.fe} \tag{8}$$

From these b equations solutions for $\widehat{\rho_j + \beta_{jh}}$ are obtained. Summing over h, solutions for the $\hat{\rho}_j$ are obtained since $\sum_{h=1}^{k_j} \hat{\beta}_{jh} = 0$. If b is less than v, then solve for $\widehat{\rho_j + \beta_{jh}}$; if not, solve for the $\hat{\tau}_i$. Of course, as a check one could obtain solutions for both sets of effects and the results must check by satisfying the normal equations. Also,

$$\hat{\mu} = \frac{1}{b}\left\{\sum_{j=1}^{r} \sum_{h=1}^{k_j} \bar{y}_{.jh} - \sum_{j=1}^{r} \sum_{h=1}^{k_j} \frac{1}{n_{.jh}} \sum_{i=1}^{v} n_{ijh}\hat{\tau}_i\right\} \\ = \frac{1}{v}\left\{\sum_{i=1}^{v} \bar{y}_{i..} - \sum_i \frac{1}{n_{i..}} \sum_{j=1}^{r} \sum_{h=1}^{k_j} n_{ijh}(\widehat{\rho_j + \beta_{jh}})\right\}. \tag{9}$$

The ρ_j^* in Table 1 are obtained from the r equations

$$\rho_j^* n_{.f.} - \sum_{i=1}^{v} \frac{n_{if.}}{n_{i..}} \sum_{j=1}^{r} n_{ij.}\rho_j^* = Y_{.f.} - \sum_{i=1}^{v} n_{if.}\bar{y}_{i..} = Q_{.f.} \tag{10}$$

plus the equation $\sum_{j=1}^{r} \rho_j^* = 0$. These equations are obtained from the normal equations for μ, τ_i, and ρ_j setting each $\beta_{jh} = 0$.

The expected value of E_e is σ_ϵ^2. The expected value of

$$\sum_{j=1}^{r} \sum_{h=1}^{k_j} (\widehat{\rho_j + \beta_{jh}})Q_{\cdot jh} = \sum_{j=1}^{r} \sum_{h=1}^{k_j} Q_{\cdot jh} \sum_{f=1}^{r} \sum_{g=1}^{k_f} k^{jhfg}Q_{\cdot fg}$$

(where k^{jhfg} are the elements of the inverse matrix in the solution of the $\widehat{\rho_j + \beta_{jh}}$ and where ρ_j and β_{jh} are random independent effects with mean zero and variances σ_ρ^2 and σ_β^2, respectively) is:

$$E\Big[E_b' = \sum_{j=1}^{r} \sum_{h=1}^{k_j} \sum_{f=1}^{r} \sum_{g=1}^{k_f} k^{jhfg}\left(Y_{.jh} - \sum_{i=1}^{v} n_{ijh}\bar{y}_{i..}\right)\left(Y_{.fg} - \sum_{d=1}^{v} n_{dfg}\bar{y}_{d..}\right)\Big]$$

$$= \sigma_\rho^2\Bigg\{\sum_{j=1}^{r} \sum_{h=1}^{k_j} k^{jhjh}\Bigg[n_{.jh}^2 - 2n_{.jh} \sum_{i=1}^{v} \frac{n_{ijh}n_{ij.}}{n_{i..}} \\ + \sum_{i=1}^{v} \frac{n_{ijh}}{n_{i..}} \sum_{d=1}^{v} \frac{n_{djh}}{n_{d..}} \sum_{e=1}^{r} n_{de.}n_{ie.}\Bigg]$$

$$\begin{aligned}
&+ \sum_{j=1}^{r} \sum_{\substack{h=1\\h\neq g}}^{k_j} \sum_{g=1}^{k_j} k^{jhjg} \Big[n_{.jh} n_{.jg} - n_{.jh} \sum_{i=1}^{v} \frac{n_{ijg} n_{ij.}}{n_{i..}} \\
&- n_{.jg} \sum_{d=1}^{v} \frac{n_{djh} n_{dj.}}{n_{d..}} + \sum_{i=1}^{v} \frac{n_{ijh}}{n_{i..}} \sum_{d=1}^{v} \frac{n_{djg}}{n_{d..}} \sum_{e=1}^{r} n_{ie.} n_{de.} \Big] \\
&- \sum_{j=1}^{r} \sum_{\substack{h=1\\jh\neq fg\\j\neq f}}^{k_j} \sum_{f=1}^{r} \sum_{g=1}^{k_f} k^{jhfg} \Big[n_{.jh} \sum_{i=1}^{r} \frac{n_{ifg} n_{ij.}}{n_{i..}} + n_{.fg} \sum_{d=1}^{v} \frac{n_{djh} n_{df.}}{n_{d..}} \\
&- \sum_{i=1}^{v} \frac{n_{ijh}}{n_{i..}} \sum_{d=1}^{v} \frac{n_{djh}}{n_{d..}} \sum_{e=1}^{r} n_{ie.} n_{de.} \Big] \Big\} \\
&+ \sigma_\beta^2 \Big\{ \sum_{j=1}^{r} \sum_{h=1}^{k_j} k^{jhjh} \Big[n_{.jh}^2 - 2n_{.jh} \sum_{i=1}^{v} \frac{n_{ijh}}{n_{i..}} \\
&+ \sum_{i=1}^{v} \frac{n_{ijh}}{n_{i..}} \sum_{d=1}^{v} \frac{n_{djh}}{n_{d..}} \sum_{e=1}^{r} \sum_{c=1}^{k_e} n_{iec} n_{dec} \Big] \\
&- \sum_{j=1}^{r} \sum_{\substack{h=1\\jh\neq fg}}^{k_j} \sum_{f=1}^{r} \sum_{g=1}^{k_f} k^{jhfg} \Big[n_{.jh} \sum_{d=1}^{v} \frac{n_{dfg} n_{djh}}{n_{d..}} + n_{.fg} \sum_{i=1}^{v} \frac{n_{ifg} n_{ijh}}{n_{i..}} \\
&- \sum_{i=1}^{v} \frac{n_{ijh}}{n_{i..}} \sum_{d=1}^{v} \frac{n_{dfg}}{n_{d..}} \sum_{e=1}^{r} \sum_{c=1}^{k_e} n_{iec} n_{dec} \Big] \Big\} \\
&+ \sigma_\epsilon^2 (b-1) = K_1 \sigma_\rho^2 + K_2 \sigma_\beta^2 + (b-1)\sigma_\epsilon^2 \, .
\end{aligned} \tag{11}$$

The expected value of $\sum_{j=1}^{r} \rho_j^* Q_{.j.}$ is [where a^{jf} are the elements of the inverse matrix obtained in the solution of the ρ_j^* from formula (10)]:

$$\begin{aligned}
&E\Big[\sum_{j=1}^{r} Q_{.j.} \sum_{f=1}^{r} a^{jf} Q_{.f.} \\
&\quad = \sum_{j=1}^{r} \Big(Y_{.j.} - \sum_{i=1}^{v} n_{ij.} \bar{y}_{i..} \Big) \sum_{f=1}^{r} a^{jf} \Big(Y_{.f.} - \sum_{i=1}^{v} n_{if.} \bar{y}_{i..} \Big) \Big] \\
&\quad = \sigma_\rho^2 \Big\{ \sum_{j=1}^{r} a^{jj} \Big(n_{.j.}^2 - 2n_{.j.} \sum_{i=1}^{v} \frac{n_{ij.}}{n_{i..}} + \sum_{i=1}^{v} \frac{n_{ij.}}{n_{i..}} \sum_{d=1}^{v} \frac{n_{dj.}}{n_{d..}} \sum_{e=1}^{r} n_{ie.} n_{de.} \Big) \\
&\qquad - \sum_{\substack{i=1\\j\neq f}}^{r} \sum_{f=1}^{r} a^{jf} \Big(n_{.j.} \sum_{i=1}^{v} \frac{n_{if.} n_{ij.}}{n_{i..}} + n_{.f.} \sum_{d=1}^{v} \frac{n_{df.} n_{dj.}}{n_{d..}} \\
&\qquad - \sum_{i=1}^{v} \frac{n_{ij.}}{n_{i..}} \sum_{d=1}^{v} \frac{n_{dj.}}{n_{d..}} \sum_{e} n_{ie.} n_{de.} \Big) \Big\} \\
&\qquad + \sigma_\beta^2 \Big\{ \sum_{j=1}^{r} a^{jj} \Big(\sum_{h=1}^{k_j} n_{.jh}^2 - 2 \sum_{h=1}^{k_j} n_{.jh} \sum_{i=1}^{v} \frac{n_{ijh}}{n_{i..}}
\end{aligned} \tag{12}$$

$$+ \sum_{i=1}^{v} \frac{n_{ij.}}{n_{i..}} \sum_{d=1}^{v} \frac{n_{dj.}}{n_{d..}} \sum_{e=1}^{r} \sum_{c=1}^{k_e} n_{iec} n_{dec} \Bigg)$$

$$- \sum_{\substack{j=1 \\ j \neq f}}^{r} \sum_{f=1}^{r} a^{jf} \Bigg(\sum_{h=1}^{k_j} n_{.jh} \sum_{i=1}^{v} \frac{n_{if.} n_{ijh}}{n_{i..}} + \sum_{g=1}^{k_f} n_{.fg} \sum_{i=1}^{v} \frac{n_{ij.} n_{ifg}}{n_{i..}}$$

$$- \sum_{i=1}^{v} \frac{n_{ij.}}{n_{i..}} \sum_{d=1}^{v} \frac{n_{df.}}{n_{d..}} \sum_{e=1}^{r} \sum_{c=1}^{k_e} n_{iec} n_{dec} \Bigg) \Bigg\}$$

$$+ \sigma_\epsilon^2(r - 1) = K_3\sigma_\rho^2 + K_4\sigma_\beta^2 + (r - 1)\sigma_\epsilon^2 \ .$$

Henderson [1953] has obtained expected values for sums of squares from non-orthogonal classifications for different situations.

Now, $\sum_{j=1}^{r} \sum_{h=1}^{k_j} \widehat{(\rho_j + \beta_{jh})} Q_{.jh} - \sum_{j=1}^{r} \rho_j^* Q_{.j.}$ is the sum of squares for incomplete blocks within complete blocks eliminating treatment effects and has the expectation:

$$(K_2 - K_4)\sigma_\beta^2 + (b - r)\sigma_\epsilon^2 \ . \tag{13}$$

By definition the coefficient of σ_ρ^2 must be zero; hence, $K_1 = K_3$; no proof of the equality is given here (see Yates, [1938]). It should be possible to simplify the coefficient for σ_β^2 since there is a relationship between the a^{jf} and the k^{jhfg}. Perhaps this should be done prior to programming for high speed computers.

The treatment mean adjusted for incomplete and complete block effects is $\hat{\mu} + \hat{\tau}_i$. Only intrablock information is utilized in obtaining the adjusted means. The variances of adjusted means and for differences between adjusted means are given above (formulae (5) and (6)).

Recovery of Intrablock Information

The sum of squares to be minimized is

$$\begin{aligned} w \sum_{i=1}^{v} \sum_{j=1}^{r} \sum_{h=1}^{k_j} n_{ijh}(Y_{ijh} - \mu - \rho_j - \tau_i - \beta_{jh})^2 \\ + w' \sum_{j=1}^{r} \sum_{h=1}^{k_j} (Y_{.jh} - n_{.jh}(\mu + \rho_j) - \sum_i n_{ijh}\tau_i)^2/n_{.jh} \end{aligned} \tag{14}$$

where the true weights are $\omega = 1/\sigma_\epsilon^2$ and $\omega'_{jh} = 1/(\sigma_\epsilon^2 + n_{.jh}\sigma_\beta^2)$ for β_{jh} independently distributed with mean zero and variance σ_β^2 and where the estimated weights are[8] $w = 1/\hat{\sigma}_\epsilon^2$ and $w'_{jh} = 1/(\hat{\sigma}_\epsilon^2 + n_{.jh}\hat{\sigma}_\beta^2)$. Instead of using a different weight for each incomplete block, an average coefficient is utilized and is given below. The estimated weight, w'_{jh} , should be utilized if there is sizeable disparity among the $n_{.jh}$. The resulting normal equations for $w'_{jh} = w'$ are:

[8] It is assumed that variable $n_{.jh}$ has no effect on the intrablock variance, (see Finney [1956]).

$$\mu\colon (w + w')\left\{\mu n_{\ldots} + \sum_{j=1}^{r} n_{.j.}\rho_j + \sum_{i=1}^{v} n_{i..}\tau_i\right\} + w \sum_{j=1}^{r}\sum_{h=1}^{k_j} n_{.jh}\beta_{jh} = (w + w')Y_{\ldots} \tag{15}$$

or

$$n_{\ldots}\mu + \sum n_{.j.}\rho_j + \sum n_{i..}\tau_i + \frac{w}{w + w'} \sum_j \sum_h n_{.jh}\beta_{jh} = Y_{\ldots}\,,$$

$$\begin{aligned}\tau_g \colon wn_{g..}\tau_g + w' \sum_{i=1}^{v} \tau_i \sum_{j=1}^{r}\sum_{h=1}^{k_j} \frac{n_{ijh}n_{gjh}}{n_{.jh}} &\\ + (w + w') \sum_{j=1}^{r} n_{gj.}(\mu + \rho_j) + w \sum_{j=1}^{r}\sum_{h=1}^{k_j} n_{gjh}\beta_{jh} &\\ = wY_{g..} + w' \sum_{j=1}^{r}\sum_{h=1}^{k_j} n_{gjh}\bar{y}_{.jh}\,, &\end{aligned} \tag{16}$$

$$\rho_j \colon (w + w')\left\{n_{.j.}(\mu + \rho_j) + \sum_{i=1}^{v} n_{ij.}\tau_i\right\} + w \sum_{h=1}^{k_j} n_{.jh}\beta_{jh} = (w + w')Y_{.j.} \tag{17}$$

or

$$n_{.j.}(\mu + \rho_j) + \sum_{i=1}^{v} n_{ij.}\tau_i + \frac{w}{w + w'} \sum_{j=1}^{k_j} n_{.jh}\beta_{jh} = Y_{.j.}\,,$$

$$\beta_{jh} \colon w\left\{n_{.jh}(\mu + \rho_j + \beta_{jh}) + \sum_{i=1}^{v} n_{ijh}\tau_i = Y_{.jh}\right\}. \tag{18}$$

Substituting for β_{jh} from equation (18) in equations (15) and (17) results in

$$\mu\colon n_{..}\mu + \sum_{j=1}^{r} n_{.j.}\rho_j + \sum_{i=1}^{v} n_{i..}\tau_i = Y_{\ldots}\,, \tag{19}$$

$$\rho_j \colon n_{.j.}(\mu + \rho_j) + \sum_{i=1}^{v} n_{ij.}\tau_i = Y_{.j.}\,. \tag{20}$$

Substituting for β_{jh} from (18) and for μ and ρ_j from (19) and (20) we obtain:

$$\begin{aligned}wn_{g..}\tau_g - (w - w') \sum_{i=1}^{v} \tau_i \sum_{j=1}^{r}\sum_{h=1}^{k_j} \frac{n_{ijh}n_{gjh}}{n_{.jh}} - w' \sum_j \frac{n_{gj.}}{n_{.j.}} \sum_i n_{ij.}\tau_i &\\ = w\left\{Y_{g..} - \sum_{j=1}^{r}\sum_{h=1}^{k_j} n_{gjh}\bar{y}_{.jh}\right\} &\\ + w'\left\{\sum_{j=1}^{r}\sum_{h=1}^{k_j} n_{gjh}\bar{y}_{.jh} - \sum_{j=1}^{r} n_{gj.}\bar{y}_{.j.}\right\} = Z_{g\ldots}\,. &\end{aligned} \tag{21}$$

The above r equations plus an additional equation, e.g., $\sum \tau_i^* = 0$, results in unique solutions for the τ_i^* ; thus

$$\begin{bmatrix} \tau_1^* \\ \tau_2^* \\ \cdots \\ \tau_v^* \end{bmatrix} = \begin{bmatrix} c^{11} & c^{12} & \cdots & c^{1v} \\ c^{21} & c^{22} & \cdots & c^{2v} \\ \cdots & \cdots & \cdots & \cdots \\ c^{v1} & c^{v2} & \cdots & c^{vv} \end{bmatrix} \begin{bmatrix} Z_{1..} \\ Z_{2..} \\ \cdots \\ Z_{v..} \end{bmatrix} \tag{22}$$

where the original equations were in the form:

$$\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1v} & 1 \\ n_{21} & n_{22} & \cdots & n_{2v} & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ n_{v1} & n_{v2} & \cdots & n_{vv} & 1 \\ 1 & 1 & & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \tau_1^* \\ \tau_2^* \\ \cdots \\ \tau_v^* \\ 0 \end{bmatrix} = \begin{bmatrix} Z_{1..} \\ Z_{2..} \\ \cdots \\ Z_{v..} \\ 0 \end{bmatrix} \tag{23}$$

and where the c^{iu} are the elements of the inverse of the matrix of coefficients for the τ_i^* .

The variances of a difference between two adjusted means recovering interblock information, say $\mu^* + \tau_1^*$ and $\mu^* + \tau_2^*$, is

$$V(\tau_1^* - \tau_2^*) = c^{11} + c^{22} - c^{12} - c^{21}, \tag{24}$$

and the variance of τ_i^* is:

$$V(\tau_i^*) = c^{ii}. \tag{25}$$

It should be noted that σ_ϵ^2 and σ_β^2 appear in the expected values of the $Z_{i..}$.

Now let us return to the calculation of the weight $w' = 1/(\hat{\sigma}_\epsilon^2 + \bar{n}\hat{\sigma}_\beta^2)$. $\bar{n}$ is determined from a nested classification in which there are no treatment effects (Federer, [1955], page 106). In the present notation the expected value of the mean square for among incomplete blocks within complete blocks in the absence of treatment effects is:

$$\sigma_\epsilon^2 + \frac{1}{b - r}\left\{n_{...} - \sum_{j=1}^{r} \sum_{h=1}^{k_j} \frac{n_{.jh}^2}{n_{.j.}}\right\}\sigma_\beta^2 \, ,$$

and hence

$$\bar{n} = \left\{n_{...} - \sum_{j=1}^{r} \sum_{h=1}^{k_j} \frac{n_{.jh}^2}{n_{.j.}}\right\}/(b - r),$$

where $b = \sum_{j=1}^{r} k_j$.

The weight w' for the jhth incomplete block will be (for individual block weights)

$$w'_{jh} = 1/(\hat{\sigma}_\epsilon^2 + n_{.jh}\hat{\sigma}_\beta^2).$$

The expected value of the mean square, E_b, for incomplete blocks within complete blocks eliminating treatment and complete block effects is:

$$\sigma_\epsilon^2 + \sigma_\beta^2(K_2 - K_4)/(b - r) = \sigma_\epsilon^2 + \bar{m}\sigma_\beta^2 .$$

The average amount of interblock information is estimated as follows:

$$w' = \bar{m}/[\bar{n}E_b - (\bar{n} - \bar{m})E_e], \tag{26}$$

and the intrablock information is estimated as

$$w = 1/E_e .$$

With these weights it is now possible to obtain solutions for the τ_i^* in (22).

Incomplete Blocks Not Arranged in Complete Blocks

The previous results may be used directly for an incomplete block design for which the b incomplete blocks are completely randomized. To apply the formulae set $\rho_j = 0$, $j = 0$, $k_j = b$, $n_{ij} = 0$ or 1, and $r = 0$. The equations then become:

$$Y_{ij} = n_{ih}(\mu + \tau_i + \beta_h + \epsilon_{ih}), \qquad Y_{..} = n_{..}\mu + \sum_{i=1}^{v} n_{i.}\tau_i + \sum_{h=1}^{b} \beta_h ,$$

$$Y_{i.} = n_{i.}(\mu + \tau_i) + \sum_{h=1}^{b} n_{ih}\beta_h , \qquad Y_{.h} = n_{.h}(\mu + \beta_h) + \sum_{i=1}^{v} n_{ih}\tau_i ,$$

$$n_{f.}\beta_f - \sum_{i=1}^{v} \frac{n_{if}}{n_{i.}} \sum_{h=1}^{b} n_{ih}\beta_h = Y_{.f} - \sum_{i=1}^{v} n_{if}\bar{y}_{i.},$$

$$n_{g.}\tau_g - \sum_{j=1}^{b} \frac{n_{gh}}{n_{.h}} \sum_{i=1}^{v} n_{ih}\tau_i = Y_{g.} - \sum_{h=1}^{b} n_{gh}\bar{y}_{.h}$$

and the expected value for blocks eliminating treatment effects sum of squares becomes:

$$\begin{aligned}(b - 1)\sigma_\epsilon^2 + \sigma_\beta^2 \Bigg\{ &\sum_{h=1}^{b} k^{hh}\left(n_{.h}^2 - 2n_{.h} \sum_i \frac{n_{ih}}{n_{i.}} + \sum_{i=1}^{v} \frac{n_{ih}}{n_{i.}} \sum_{d=1}^{v} \frac{n_{dh}}{n_{d.}} \sum_{e=1}^{b} n_{ie}n_{de}\right) \\ &- \sum_{\substack{h=1 \\ h\neq f}}^{b} \sum_{f=1}^{b} k^{hf}\Bigg(n_{.h} \sum_{d=1}^{v} \frac{n_{df}n_{dh}}{n_{d.}} + n_{.f} \sum_{i=1}^{v} \frac{n_{if}n_{ih}}{n_{i.}} \\ &- \sum_{i=1}^{v} \frac{n_{ih}}{n_{i.}} \sum_{d} \frac{n_{df}}{n_{d.}} \sum_{e=1}^{b} n_{ie}n_{de}\Bigg)\Bigg\} = (b - 1)\sigma_\epsilon^2 + (b - 1)\bar{m}\sigma_\beta^2 .\end{aligned} \tag{27}$$

Utilizing these results, the analysis goes through in much the same manner as for the experimental design in which the incomplete blocks are arranged in complete blocks.

If the information from complete blocks as well as from incomplete blocks within complete blocks is utilized, the sum of squares is $\sum_{j=1}^{r} \sum_{h=1}^{k_j} \widehat{(\rho_j + \beta_{jh})} Q_{.jh}$ and its expectation is obtained from equation (11) which reduces to (27) above. If the differences among complete blocks are random effects, perhaps the weight w' should be computed from E_b' instead of E_b in the analysis of variance table.

REFERENCES

Basson, R. P. [1959]. *Incomplete Block Designs Augmented with a Repeated Control.* M.S. Thesis, Iowa State Univ. Library.

Bose, R. C. and Nair, K. R. [1939]. Partially balanced incomplete block designs. *Sankhyā 4*, 337–72.

Corsten, L. C. A. [1959]. *Incomplete Block Designs in Which the Number of Replicates is not the Same for All Treatments.* Inst. of Statistics Mimeo. Series No. 226, Univ. N. C.

Das, M. N. [1958]. On reinforced incomplete block designs. *Jour. Indian Soc. Agric. Stat. 10*, 73–7.

Federer, W. T. [1955]. *Experimental Design—Theory and Application.* Macmillan, N. Y.

Federer, W. T. [1956a]. *Augmented (or hoonuiaku) designs.* Biometrics Unit, Cornell Univ. Mimeo. BU-74-M, February.

Federer, W. T. [1956b]. A method for evaluating genetic progress in a sugar cane breeding program. *Hawaiian Planters' Record 55*, 177–90.

Federer, W. T. [1956c]. Augmented (or hoonuiaku) designs. *Hawaiian Planters' Record 55*, 191–208.

Federer, W. T. [1957]. Variance and covariance analyses for unbalanced classifications. *Biometrics 13*, 333–62.

Federer, W. T. [1958]. Augmented designs (abstract). *Biometrics 14*, 134.

Finney, D. J. [1956]. The statistician and the planning of field experiments. *J. Roy. Stat. Soc. Series A*, 119, 1–27.

Graybill, F. A. and Pruitt, W. E. [1958]. The staircase design: theory. *Annals of Math. Stat. 29*, 523–33.

Henderson, C. R. [1953]. Estimation of variance and covariance components. *Biometrics 9*, 226–52.

Justensen, S. H. and Keuls, M. [1958]. Note on the use of non-orthogonal designs. *ISI Bulletin 36*, 269–76.

Kempthorne, O. [1952]. *The Design and Analysis of Experiments.* Wiley, N. Y.

Kishen, K. [1941]. Symmetrical unequal block arrangements. *Sankhyā 5*, 329–44.

McIntyre, G. A. [1958]. Designs with unequal numbers of replications. *Unpublished paper.*

Nair, K. R. [1941]. A note on the method of "fitting of constants" for analysis of non-orthogonal data arranged in a double classification. *Sankhyā 5*, 317–28.

Pearce, S. C. [1948]. Randomized blocks with interchanged and substituted plots. *J. Roy. Stat. Soc. Series B, 10*, 252–56.

Rao, C. R. [1947]. General methods of analysis for incomplete block designs. *J. Amer. Stat. Assoc. 42*, 541–61.

Yates, F. [1934]. The analysis of multiple classifications with unequal numbers in the different classes. *J. Amer. Stat. Assoc. 29*, 51–66.

Yates, F. [1936a]. Incomplete randomized blocks. *Ann. Eug. 7*, 121–40.

Yates, F. [1936b]. A new method of arranging variety trials involving a large number of varieties. *J. Agric. Sci. 26*, 424–55.

Yates, F. [1938]. Orthogonal functions and tests of significance in the analysis of variance. *J. Roy. Stat. Soc. Series B, 5*, 177–80.

Youden, W. J. and Connor, W. S. [1953]. The chain block design. *Biometrics 9*, 127–40.

PHENOTYPIC, GENETIC AND ENVIRONMENTAL CORRELATIONS

S. R. Searle
N. Z. Dairy Board, Wellington, New Zealand.

INTRODUCTION

A phenotypic correlation is the correlation between records of two traits on the same animal and is usually estimated by the product-moment correlation statistic. The genetic correlation, on the other hand, is the correlation between an animal's genetic value for one trait and the same animal's genetic value for the other trait, estimators for which have been proposed by Hazel [1943]. Estimates of these correlations are widespread throughout the literature of animal breeding and in many instances the estimate of a phenotypic correlation is reported smaller in magnitude than that of the corresponding genetic correlation, e.g. with certain poultry records, in Lerner & Cruden [1948], sheep records in Morley [1951] and with certain dairy records in VanVleck [1960] and Searle [1961]. Such results may seem a little unexpected at first sight since phenotype includes genotype and one might anticipate the correlation between phenotypes to be larger than that between genotypes. When estimates have not followed this pattern the explanation is sometimes given that a phenotypic correlation less than a genetic correlation is the result of a negative environmental correlation in the records of the two traits. This paper investigates the relationship between these three correlations on the basis of a linear model, and demonstrates the situations in which this explanation is correct. Other comparisons are also made.

LINEAR MODELS

Suppose the records of two traits in an animal are x and X. Neglecting the general means, we will take each variable as being the sum of a genetic term and an environmental (including error) term, i.e.

$$x = g + e,$$

and

$$X = G + E. \tag{1}$$

The genetic correlation r is the correlation between g and G; that between x and X is the phenotypic correlation, R say, and we will define

r' as the environmental correlation, namely that between e and E. Thus we have

$$r = \text{cov}\,(g, G)/\sigma_g\sigma_G \,,$$

$$R = \text{cov}\,(x, X)/\sigma_x\sigma_X \,,$$

and

$$r' = \text{cov}\,(e, E)/\sigma_e\sigma_E \,.$$

The covariance between g and G is denoted by cov (g, G) and their variances are σ_g^2 and σ_G^2 respectively, with similar notation for the other terms. The phenotypic correlation R is that between x and X which can be obtained directly from (1) as

$$R = \frac{\text{cov}\,(g, G) + \text{cov}\,(g, E) + \text{cov}\,(G, e) + \text{cov}\,(e, E)}{\sqrt{[\sigma_g^2 + 2\,\text{cov}\,(g, e) + \sigma_e^2][\sigma_G^2 + 2\,\text{cov}\,(G, E) + \sigma_E^2]}}\,.$$

Assuming all genetic-environment covariances (i.e. interactions) are zero this becomes

$$R = \frac{\text{cov}\,(g, G) + \text{cov}\,(e, E)}{\sqrt{(\sigma_g^2 + \sigma_e^2)(\sigma_G^2 + \sigma_E^2)}}\,.$$

This reduces to

$$R = r\sqrt{hH} + r'\sqrt{(1-h)(1-H)} \qquad (3)$$

where h and H are the heritabilities in the narrow sense, of the two traits, defined as $\sigma_g^2/(\sigma_g^2 + \sigma_e^2)$ and $\sigma_G^2/(\sigma_G^2 + \sigma_E^2)$ respectively. This relationship between the three correlations, phenotypic, genetic and environmental, is derived in Lerner [1950] using the method of path coefficients. In both cases genetic-environmental interactions have been assumed zero, as is customary in discussions of this nature.

ENVIRONMENTAL CORRELATION

Equation (3) can be re-arranged to give the environmental correlation as

$$r' = (R - r\sqrt{hH})/\sqrt{(1-h)(1-H)}. \qquad (4)$$

In terms of the model (1) this is the correlation between e and E, which include the random errors. Assuming the correlation between these is zero, r' can be thought of as the environmental correlation. Phenotypic correlations are usually estimated directly whereas genetic correlations are derived from covariance analyses between relatives and contain additive genetic variation only. Environmental correlations estimated from this formula may therefore contain genetic elements over and above additive genetic variation.

Consideration of (4) shows that when R and r have the same sign r' will be negative if r is greater than $R/\sqrt{hH}$. When r and R are of opposite sign r' has the same sign as R. Thus the phenotypic and genetic correlations being of opposite sign (an infrequent occurrence one would imagine) implies that the phenotypic and environmental correlations have the same sign.

Genetic and phenotypic correlations of similar sign is the usual situation, and in this case we see that the ratio of the phenotypic to the genetic correlation has to be less than the geometric mean of the heritabilities, before the phenotypic correlation being less than the genetic correlation implies a negative environmental correlation. Values of this geometric mean are given in Table 1.

TABLE 1

VALUES OF $A = \sqrt{hH}$[1]

Heritability of one trait	Heritability of second trait									
	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	.10									
.2	.14	.20								
.3	.17	.24	.30							
.4	.20	.28	.35	.40						
.5	.22	.32	.39	.45	.50					
.6	.24	.35	.42	.49	.55	.60				
.7	.26	.37	.46	.53	.59	.65	.70			
.8	.28	.40	.49	.57	.63	.69	.75	.80		
.9	.30	.42	.52	.60	.67	.73	.79	.85	.90	
1.0	.32	.45	.55	.63	.71	.77	.84	.89	.95	1.00

[1] The environmental correlation is negative when the ratio of the phenotypic correlation to the genetic correlation is less than A.

This shows that for traits with low heritabilities (as is the case with many traits of economic importance in farm animals) the ratio

phenotypic correlation/genetic correlation

has also to be low before a negative environmental correlation is implied. For example with heritabilities 0.5 and 0.3 this ratio must be less than 0.39, i.e. the genetic correlation must be more than two-and-a-half times as great as the phenotypic correlation for the environmental correlation to be negative.

A phenotypic correlation less than its genetic counterpart, together with a small positive environmental correlation, will occur where the genes governing two traits are similar but where the environments pertaining to the expression of these traits have a low correlation. For example, the genes controlling milk production in the first month of a cow's lactation and those controlling total lactation yield may be quite highly correlated; but the environments pertaining to the first month and to the lactation yields may have a low correlation. The phenotypic correlation would then be less than the genetic correlation. This is observed in estimates reported by Searle [1961] and VanVleck [1960] and also by Lerner and Cruden [1948] for egg production. The situation of an environmental correlation having a small value is by no means universal and in many instances its value will be large (positive or negative) because the environment generally affects an individual in all its parts and functions.

The equation (3) for the phenotypic correlation R can be written as

$$R = Ar + Br' \tag{5}$$

where $A = \sqrt{hH}$ and $B = \sqrt{(1 - h)(1 - H)}$. Because $A < 1$ a negative value of r' implies R being less than r. Thus a negative environmental correlation always implies the phenotypic correlation being less than the genetic correlation, although it can be less without the environmental correlation necessarily being negative.

The equation for R when the heritabilities are equal is

$$R = hr + (1 - h)r'$$

from which it is seen that if any two of R, r and r' are equal the third is also. Thus equal heritabilities imply that equality of any two of the correlations is tantamount to equality of all three.

PHENOTYPIC AND GENETIC CORRELATIONS.

The relationship of R and r' to r can be discussed in terms of equation (5), first noting that $A + B < 1$ because $A < \frac{1}{2}(h + H)$ and $B < \frac{1}{2}(2 - h + H)$ arising from a property of geometric and arithmetic means. The phenotypic correlation R, can only exceed the genetic, r, when the environmental correlation r' also exceeds r, and sufficiently so; in fact r' must be greater than $r(1 - A)/B$. Thus $R \gtrless r$ according as $r'/r \gtrless (1 - A)/B$, and since $(1 - A)/B > 1$ there is a small region where $R < r$ but $r' > r$; otherwise R and r' are less than r together. Values of $(1 - A)/B$ are shown in Table 2. An example is for heritabilities of 0.3 and 0.5, for which the entry in the Table is 1.03; thus, if the environmental correlation is less than 1.03 times the genetic correlation,

TABLE 2

VALUES OF $(1 - A)/B = \dfrac{1 - \sqrt{hH}}{\sqrt{(1 - h)(1 - H)}}$[2]

Heritability of one trait	Heritability of second trait									
	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.1	1.00									
.2	1.01	1.00								
.3	1.05	1.01	1.00							
.4	1.10	1.04	1.00	1.00						
.5	1.16	1.08	1.03	1.00	1.00					
.6	1.27	1.14	1.09	1.04	1.00	1.00				
.7	1.42	1.29	1.17	1.12	1.05	1.00	1.00			
.8	1.71	1.50	1.38	1.23	1.16	1.11	1.00	1.00		
.9	2.33	2.07	1.85	1.67	1.50	1.35	1.24	1.07	1.00	

[2] The genetic correlation exceeds the phenotypic correlation when the ratio of the environmental correlation to the genetic correlation is less than $(1\text{-}A)/B$.

then the phenotypic correlation is less than the genetic correlation. The values in this Table are close to 1.00 for heritabilities that are small or alike, and for practical purposes in these cases both r' and R exceed or are less than r together. But when the heritabilities are unequal the fraction is larger, for example, its value is 1.50 when $h = 0.2$ and $H = 0.8$.

The equation for R represents a plane in the 3-dimensional system having co-ordinate axes for R, r and r', but the relationships just discussed can be illustrated by plotting the line of intersection of the planes $R = Ar + Br'$ and $R = k$ on the plane $R = 0$, for various values of k. This is shown in Figure 1 for the example $h = 0.2$ and $H = 0.8$ with $R = 0.4r'$, and $R < r$ for $r' < 1.50r$. Lines for $R = 0.3$ and $R = 0.2$ are shown intersected by $r' = 1.50r$. Below this $R < r$ and above it $R > r$. Between $r' = 1.50r$ and $r' = r$, R is less than r and r' exceeds r; i.e. the phenotypic correlation is less than the genetic but the environmental exceeds it. Only the first quadrant of the (r', r) Cartesian system is shown, but extension of the lines for R into the second and fourth quadrants demonstrates the conditions under which the genetic and phenotypic correlations are of different sign, and the environmental correlation is negative.

SUMMARY

(i) The phenotypic, genetic and environmental correlations, R, r

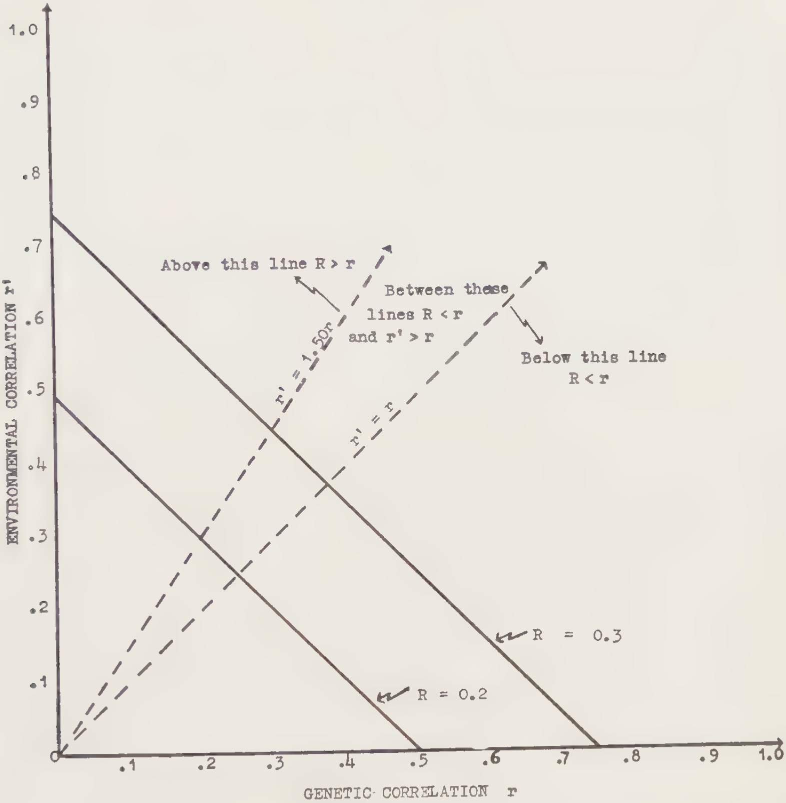


FIGURE 1

PHENOTYPIC CORRELATIONS $R = 0.2$ AND 0.3 FOR HERITABILITIES 0.2 AND 0.8.

and r', are connected by the relationship

$$R = r\sqrt{hH} + r'\sqrt{(1-h)(1-H)}$$

h and H being the heritabilities of the traits involved.

(ii) The environmental correlation is negative when R and r have the same sign only if $R/r < \sqrt{hH}$; it is negative when R and r are of opposite sign and R is negative.

(iii) Equality of the heritabilities implies that when any two of the correlations are equal there is equality of all three.

(iv) The phenotypic correlation exceeds (or is less than) the genetic correlation according as the ratio of the environmental to the genetic correlation exceeds (or is less than) the value of $(1 - \sqrt{hH})/\sqrt{(1-h)(1-H)}$.

REFERENCES

1. Hazel L. N. [1943]. The genetic basis for constructing selection indexes. *Genetics 28*, 476–90.
2. Lerner I. M. [1950]. Population Genetics and Animal Improvement. Cambridge University Press.
3. Lerner I. M. & Cruden D. M. [1948]. The heritability of accumulative monthly and annual egg production. *Poultry Sci. 27*, 67–78.
4. Morley F. H. W. [1951]. Selection for economic characters in Australian Merino sheep. I. Estimates of phenotypic and genetic parameters. *Science Bull. No. 73*, N.S.W. Dept. Agric.
5. Searle S. R. [1961]. Part Lactations II. Genetic and phenotypic studies of monthly milk fat yields. *J. Dairy Sci. 44*, 282–295.
6. VanVleck L. D. [1960]. *The Value of Part Lactation Records in Selection.* Unpublished Ph.D. Thesis, Cornell University Library, Ithaca, N. Y.

RELATIVE EFFICIENCIES OF HERITABILITY ESTIMATES BASED ON REGRESSION OF OFFSPRING ON PARENT[1]

B. B. Bohren, H. E. McKean,

Population Genetics Institute, Purdue University, Lafayette, Indiana, U.S.A.

AND

Yukio Yamada

National Institute of Genetics, Misima, Japan

INTRODUCTION

The problem of optimal estimation of the coefficient of regression of offspring on parent (in the sense of minimum variance), when the number of progeny per parent is arbitrary, was completely solved by Kempthorne and Tandon [1953]. Prior to this paper, two methods were (and still are) commonly used: (1) the regression of the phenotypic mean of all offspring of a given parent on the parent's record; (2) the regression of offspring on parent, in which the parent's record is repeated for each of its progeny. Kempthorne and Tandon's technique, which they refer to as (3) the weighted regression technique, assigns weights to the progeny means which are functions of the number of progeny and a guessed value of a correlation coefficient ρ between deviations from regression associated with two progeny of the same parent. The difficulty here lies in the fact that ρ is unknown. The success of the general technique depends upon guessing ρ accurately. Presumably if the guessed value of ρ is close to ρ, the weighted technique is close to optimal. The precise effect of a poor guess for ρ does not seem to be known.

The purposes of this paper are (1) to investigate the nature and magnitude of the correlation coefficient ρ, and (2) to compare the efficiencies of the various techniques with respect to data from a population of poultry.

THEORY

General: Following the notation of Kempthorne and Tandon, the general model for the regression of offspring on one parent would be

[1] Journal Paper Number *1659*, Purdue University Agricultural Experiment Station.

$$Y_{jk} = \mu_Y + \beta(X_j - \mu) + e_{jk} , \qquad (1)$$

where

Y_{jk} = the phenotypic value of the kth progeny of parent j,
X_j = the phenotypic value of parent j,
μ_Y = average phenotypic value of the offspring population,
μ = average phenotypic value of the parent population,
e_{jk} = the deviation peculiar to the kth progeny of parent j,

and

β = the regression coefficient of Y on X.

From such a model Kempthorne and Tandon showed that

$$E(e_{jk}^2) = \sigma_P^2(1 - \beta^2),$$

$$E(e_{jk}e_{jk'}) = \sigma_P^2(\rho_1 - \beta^2),$$

and

$$\rho = (\rho_1 - \beta^2)/(1 - \beta^2),$$

where ρ_1 is the correlation between progeny having a common parent (X_j). This result is completely general when σ_P^2 = Var (Y_{jk}) = Var (X_j) and would apply to situations in which non-additivity existed in the underlying genetic model and/or if correlations existed between the environments within progeny groups. In such a case, ρ_1 could be considered as the repeatability between members of a progeny group.

In order to simplify the considerations to follow, a base population having certain characteristics is assumed.

(i) Only additive genetic variance is present.
(ii) Effects of environment are random, so that the environmental correlation between individuals in a progeny group is essentially zero.
(iii) Random mating is practiced.

In the classical linear regression model, the e_{jk} form a set of mutually uncorrelated random variables. That such is not the case in this instance can be clearly demonstrated, even in the completely additive situation. If we knew the true breeding values (genotypic values in the additive genetic model) of both the sires and the dams, and mating was random, we would use the model

$$Y_{ijk} = \mu_Y + \tfrac{1}{2}(g_j - \mu) + \tfrac{1}{2}(g_i - \mu) + \delta_{ijk} + \epsilon_{ijk} , \qquad (2)$$

where

g_j = the true breeding value of the parent of interest,

$g_{\mathbf{i}}$ = the true breeding value of the other parent,
δ_{ijk} = the individual genetic deviation from the mean of the sire and dam due to segregation,
ϵ_{ijk} = the random environmental deviation,

and

μ $= E(g_i) = E(g_{\mathbf{i}})$.

Under the assumption (i) of additive genetic variance, the total genetic variance σ_G^2 is entirely additive; hence it is clear that

$$\text{Var}\,(g_i) = \text{Var}\,(g_{\mathbf{i}}) = \sigma_G^2\,,$$

$$\text{Var}\,(\delta_{ijk}) = \sigma_G^2/2,$$

and

$$\text{Var}\,(\epsilon_{ijk}) = \sigma_E^2\,.$$

This model (2) meets the conditions of the classical model in that all random components are uncorrelated. In the application of this model, however, only the phenotypic value of one parent is known, while the progeny phenotypes center on the parent's breeding value. To illustrate, consider that with purely additive gene effects, h^2 may be considered as the regression of the parent's breeding value (in this case genotypic value) on the parent's phenotypic value, so that the estimate of the breeding value of the parent based on its phenotype is

$$(\hat{g}_i - \mu) = h^2(X_i - \mu), \tag{3}$$

but the true breeding value of the parent would be

$$(g_i - \mu) = h^2(X_i - \mu) + I_i\,, \tag{4}$$

where g_i = the breeding value of the parent, X_i = the phenotypic value of the parent, and I_i = the discrepancy between the estimated and actual breeding values. It is important to note that the progeny of the parent will vary around the point determined by (4) and not around the point determined by (3). It is this fact which causes the correlation between the errors of progeny from the same parent.

Therefore, in applying equation (2) we must replace $(g_i - \mu)$ by (4), yielding, since $\beta = h^2/2$ under the assumptions,

$$Y_{ijk} = \mu_Y + \beta(X_i - \mu) + \tfrac{1}{2}I_i + \tfrac{1}{2}(g_{\mathbf{i}} - \mu) + \delta_{ijk} + \epsilon_{ijk}\,. \tag{5}$$

Comparing (1) and (5) we see that the term in (5) comparable to e_{jk} in (1) is

$$e_{ijk} = \tfrac{1}{2}I_i + \tfrac{1}{2}(g_{\mathbf{i}} - \mu) + \delta_{ijk} + \epsilon_{ijk}\,. \tag{6}$$

From (5) and (6) is obtained the model

$$Y_{ijk} = \mu_y + \beta(X_i - \mu) + e_{ijk} \,, \tag{5'}$$

which is virtually identical to (1). The subscript i in (5′) is actually superflous, since the effect of the sire is embedded in e_{ijk}, so that dropping the subscript i yields model (1).

Since $E(e_{ijk}) = 0$, it follows from either (6) or (4) that $E(I_i) = 0$. Also, the variance of I_i would be

$$E(I_i^2) = (1 - h^2)\sigma_G^2 \,.$$

Special Cases: *I*. If we now make the further simplifying assumption (iv) that no two progeny of the given parent (the dam) have the same parent of the opposite sex (sire), we may rewrite (5) as,

$$Y_{jk} = \mu_Y + \beta(X_i - \mu) + \tfrac{1}{2}I_i + \tfrac{1}{2}(g_{ik} - \mu) + \delta_{jk} + \epsilon_{jk} \,. \tag{7}$$

Now from (6) or (7) we see that,

$$E(e_{jk}^2) = \tfrac{1}{4}(1 - h^2)\sigma_G^2 + \tfrac{1}{4}\sigma_G^2 + \tfrac{1}{2}\sigma_G^2 + \sigma_E^2 = (1 - h^4/4)\sigma_P^2 \,,$$

and

$$E(e_{jk}e_{jk'}) = \tfrac{1}{4}E(I_i) = \tfrac{1}{4}(1 - h^2)\sigma_G^2 \,.$$

Then

$$\begin{aligned}\rho &= [\mathrm{Cov}\,(e_{jk}\,, e_{jk'})]/\sqrt{\mathrm{Var}\,(e_{jk})\ \mathrm{Var}\,(e_{jk'})}\\ &= [\tfrac{1}{4}(1 - h^2)\sigma_G^2]/(1 - h^4/4)\sigma_P^2 = (h^2 - h^4)/(4 - h^4).\end{aligned}$$

This result is exactly analogous to the result of Kempthorne and Tandon, for under the assumption (i)–(iv), their $\rho_1 = h^4/4$ and $\beta = h^2/2$.

II. It often happens, especially in poultry breeding operations, that matings are made up so that a sample of s sires are each mated to a distinct random sample of dams so that d_i dams are mated to sire i. In order for our assumptions to meet this situation it is only necessary to replace assumption (iv) by a new assumption to the effect that, (v) all progeny of a given parent of interest (the dam) have the same parent of the opposite sex (sire). We may now consider Y as a function of the breeding values of the sire and dam, assuming completely additive gene effects, such that

$$Y_{ijk} = \mu_Y + \tfrac{1}{2}(g_i - \mu) + \tfrac{1}{2}(g_{ij} - \mu) + \delta_{ijk} + \epsilon_{ijk} \,, \tag{8}$$

which is analogous to (2). Clearly the expected progeny mean for a given sire depends upon the breeding value of the sire, or

$$E(Y_{ijk} \mid i) = \mu_Y + \tfrac{1}{2}(g_i - \mu) = \mu_i \,. \tag{9}$$

It is always the case, however, that only the dam's phenotype

(X_{ij}) is known, hence from (8) and (4) we may write

$$Y_{ijk} = \mu_i + \frac{h^2}{2}(X_{ij} - \mu) + \tfrac{1}{2}I_{ij} + \delta_{ijk} + \epsilon_{ijk} \;, \tag{10}$$

or

$$Y_{ijk} = \mu_i + \beta(X_{ij} - \mu) + e'_{ijk} \;, \tag{11}$$

where

$$e'_{ijk} = \tfrac{1}{2}I_{ij} + \delta_{ijk} + \epsilon_{ijk} \;. \tag{12}$$

This clearly shows that the genetic interpretation of the coefficient of regression of offspring phenotypic values on the dam's phenotypic value is the same, regardless of whether the dams are mated to a random group of sires or to a single sire. On the intra-sire basis, β is estimated within sires, one estimate for each sire, and a linear combination of these estimates is taken as a single point estimate of β. From (11) we see that for a given sire i,

$$\begin{aligned} E(e'_{ijk} \mid i) &= 0, \\ E(e'^2_{ijk} \mid i) &= \tfrac{1}{4}E(I^2_{jk}) + E(\delta^2_{ijk}) + E(\epsilon^2_{ijk}) \\ &= \sigma^2_P(4 - h^2 - h^4)/4. \end{aligned}$$

This variance is less than the variance of e_{jk} from model (7) because e'_{ijk} does not contain an effect due to sire. The covariance is the same under both assumptions since from (12) we see that

$$E(e'_{ijk}e'_{ijk'} \mid i) = \sigma^2_G(1 - h^2)/4.$$

The correlation ρ^* between the errors of progeny of a given dam for a fixed sire would then be

$$\rho^* = [\sigma^2_G(1 - h^2)/4]/[\sigma^2_P(4 - h^2 - h^4)/4] = (h^2 - h^4)/(4 - h^2 - h^4).$$

This result is again equivalent to that obtained by Kempthorne and Tandon, since

$$\rho^* = (\rho^*_1 - \beta^{*2})/(1 - \beta^{*2}),$$

where, under the assumptions including (v), ρ^*_1 is the correlation between progeny of the same dam in the conditional population of progeny of a given sire, and equals $h^2/(4 - h^2)$, while β^* would have to be the correlation between parent and offspring in the conditional population or $h^2/\sqrt{4 - h^2}$, since σ^2_Y, being the conditional variance within sires, is less than σ^2_X. For comparison, ρ^* can also be evaluated in terms of the parameters under the assumptions including (iv), in which case,

$$\rho^* = (\rho_1 - \beta^2)/(1 - \rho_1 - \beta^2).$$

Magnitude of ρ and ρ^*: The quantity ρ or ρ^*, depending on the mating structure, is the intrinsic parameter upon which the Kempthorne and Tandon weighted regression technique is based. Therefore, it seems important to consider the possible values which ρ or ρ^* may assume, in order to be able to choose the most efficient method of estimating the regression of interest, or, if the weighted technique is used, to enable an enlightened "guess" as to its value.

Examination of the values of ρ and ρ^* in terms of h^2 shows that $\rho = \rho^* = 0$ if, and only if, $h^2 = 0$ or 1. Conversely ρ or $\rho^* > 0$ for $0 < h^2 < 1$. Elementary analytical techniques show that ρ reaches a maximum value when $h^2 = .536$, at which value $\rho = .067$. Similarly, ρ^* achieves an absolute maximum when $h^2 = .586$, when $\rho^* = .079$. The functional relationship between ρ and h^2 is illustrated in Figure 1. It is recognized that these may be minimal values for ρ and ρ^*, since failure of assumption (ii) in particular will inflate these correlations. The effect on ρ of relaxing assumption (i) is not clear. The important

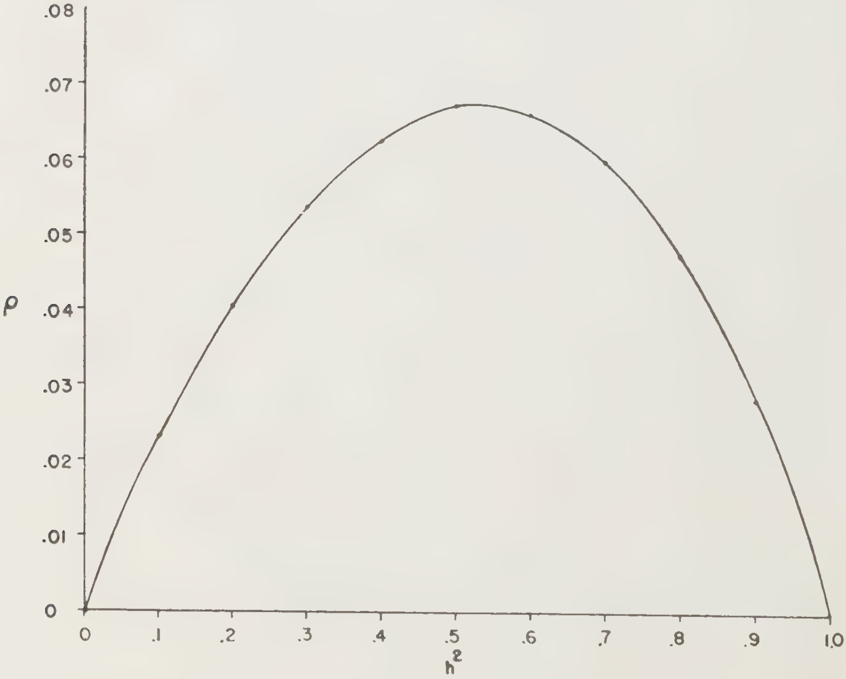


FIGURE 1

THE RELATIONSHIP BETWEEN ρ AND h^2. $\rho = (h^2 - h^4)/(4 - h^4)$

point is that these values would be unlikely to be large especially in a poultry population, where environmental correlations would be small or non-existent. In some larger species such as dairy cattle, such environmental correlations could have sizable values.

A primary advantage of the offspring on parent regression technique is that unbiased estimates of h^2 may be obtained when the parents are selected for the trait under consideration. It is of interest to determine the effect of selection of the parents on the correlation between the errors and indirectly on the variance of the estimated β's. It has already been observed that a general expression for ρ under any mating structure would be

$$\rho = (\rho_1 - r^2)/(1 - r^2) = [\rho_1 - \beta^2(\sigma_X^2/\sigma_Y^2)]/[1 - \beta^2(\sigma_X^2/\sigma_Y^2)],$$

where ρ_1 is the phenotypic correlation between progeny of the same parent and r is the correlation between parent and offspring, β being a constant. It is observed that for special Cases I and II,

$$\begin{aligned}\rho_1 &= [\mathrm{Cov}\,(Y_{jk}\,,\,Y_{jk'})]/\mathrm{Var}\,(Y_{jk}) = [\beta\sigma_X^2 + \tfrac{1}{4}E(I_{jk})^2]/\sigma_Y^2\\ &= \beta^2(\sigma_X^2/\sigma_Y^2) + \tfrac{1}{4}(1 - h^2)(\sigma_G^2/\sigma_Y^2).\end{aligned}$$

Substituting this value for ρ_1 in the preceding equation yields

$$\rho = [\tfrac{1}{4}(1 - h^2)(\sigma_G^2/\sigma_Y^2)]/[1 - \beta^2(\sigma_X^2/\sigma_Y^2)] = \tfrac{1}{4}(1 - h^2)\sigma_G^2/(\sigma_Y^2 - \beta^2\sigma_X^2).$$

It is clear that as the usual uni-directional selection occurs on the parents, the values of both the numerator and denominator will decrease. To evaluate this change it is now necessary to evaluate σ_Y^2 for a specific mating structure. For illustration assume a hierarchal structure in which

$$\sigma_Y^2 = \beta^2\sigma_X^2 + \tfrac{1}{4}(1 - h^2)\sigma_G^2 + \tfrac{1}{2}\sigma_G^2 + \sigma_E^2\,.$$

Substitution of this value for σ_Y^2 in the preceding equation yields

$$\rho^* = \frac{\tfrac{1}{4}(1 - h^2)\sigma_G^2}{\beta^2\sigma_X^2 + \tfrac{1}{4}(1 - h^2)\sigma_G^2 + \tfrac{1}{2}\sigma_G^2 + \sigma_E^2 - \beta^2\sigma_X^2} = \frac{h^2 - h^4}{4 - h^2 - h^4}.$$

It is seen that while σ_X^2 may be much smaller than σ_Y^2, due to selection of parents, the value ρ^* is independent of the value of σ_X^2. Similarly, in special Case I where each progeny of the dam has a different sire, it can be shown that the value of ρ is unaffected by selection of dams.

Comparison of the three estimation procedures: All three procedures for estimating β are of course unbiased. The variances of the estimates are the prime consideration. For any linear unbiased estimator of β, in which

$$\hat{\beta} = \sum_j w_j(x_j - \bar{x})\bar{y}_{j\cdot}/\sum_j w_j(x_j - \bar{x})^2, \tag{13}$$

the variance will be

$$\sigma^2_{\hat{\beta}} = \sigma^2(1 - \rho)[\sum_j w_j^2(x_j - \bar{x})^2(1 + n_jT)/n_j]/[\sum_j w_j(x_j - \bar{x})^2]^2, \tag{14}$$

where

$$\bar{x} = \sum_j w_jx_j/\sum_j w_j \ , \qquad T = \rho/(1 - \rho),$$

$$\bar{y}_{j\cdot} = \sum_k y_{jk}/n_j \ , \qquad \sigma^2 = \mathrm{Var}\ (e_{jk}),$$

and

w_j = a weighting factor applied to the information from the jth dam.

If a hierarchal mating structure is used, ρ would be replaced by ρ^* and T by $T^* = \rho^*/(1 - \rho^*)$ in this discussion.

Kempthorne and Tandon showed that the minimum variance resulted when

$$w_j = n_j/(1 + n_jT),$$

in which case the variance (14) reduces to

$$\sigma^2_{\hat{\beta}_3} = \sigma^2(1 - \rho)/\sum_j w_j(x_j - \bar{x})^2, \tag{15}$$

and

$$\bar{x} = \sum_j \left(\frac{n_j}{1 + n_jT}\right)x_j \Big/ \sum_j \left(\frac{n_j}{1 + n_jT}\right).$$

This is the optimum weighting applied to the Kempthorne-Tandon weighted regression technique or method (3). However, the value of T is never known exactly so that it is necessary to guess a value for T, which may be indicated by τ. The weights then are

$$w_j = n_j/(1 + n_j\tau),$$

and the variance (14) reduces to

$$\sigma^2_{\hat{\beta}_3'} = \sigma^2(1 - \rho) \sum_j n_j \frac{1 + n_jT}{(1 + n_j\tau)^2} \frac{(x_j - \bar{x})^2}{\sum_j [w_j(x_j - \bar{x})^2]^2}\ , \tag{16}$$

where

$$\bar{x} = \sum_j \left(\frac{n_j}{1 + n_j\tau}\right)x_j \Big/ \sum_j \left(\frac{n_j}{1 + n_j\tau}\right).$$

This value is larger than (15) since the weighting is not optimum, but will approach the minimum variance (15) as τ approaches T.

Method (2), or the technique in which the dam's record is repeated for each progeny record, is in fact a special case of the Kempthorne-Tandon technique in which τ is set equal to zero. In this case $w_i = n_i$ and when these weights are used in (14) or τ considered to be zero in (16) the variance becomes

$$\sigma^2_{\hat{\beta}_2} = \sigma^2(1 - \rho) \sum_i n_i(1 + n_i T)(x_i - \bar{\bar{x}})^2/[\sum_i n_i(x_i - \bar{\bar{x}})^2]^2, \qquad (17)$$

in which $\bar{\bar{x}} = \sum_i n_i x_i / \sum_i n_i$.

In method (1), in which the progeny means are regressed on the dam's record, w_i is simply set equal to one. From (14) the variance of this estimate is

$$\sigma^2_{\hat{\beta}_1} = \sigma^2(1 - \rho) \sum_i \frac{(1 + n_i T)}{n_i} (x_i - \bar{x})^2/[\sum_i (x_i - \bar{x})^2]^2, \qquad (18)$$

and $\bar{x} = \sum_i x_i/d$, where d is the number of dams.

It may be noted that if $n_1 = n_2 = \cdots n_p = n$, then (14) reduces to

$$\sigma^2_{\hat{\beta}} = \sigma^2[1 + (n - 1)\rho]/n \sum_i (x_i - \bar{x})^2, \qquad (19)$$

and is the same for all three estimation techniques.

The minimum variance occurs when $T = \tau$. Consequently if $T = 0$, then the repeated dam technique (method 2) will yield the minimum variance. If $T \neq 0$, then the weighted technique (method 3) will approach the minimum variance provided the estimate τ is close to T. Since the relationship between h^2 and ρ is clear (granted the genetic assumptions), an intelligent guess for τ can usually be made by applying prior knowledge.

The variances of the three estimates of β are also considerably affected by the distribution of the n_i values. A detailed discussion of this point will be presented in a subsequent paper.

AN ILLUSTRATION

Kempthorne and Tandon found little difference between the variances of the three estimates in their illustration involving dairy cattle data. As they point out, this result is not surprising since the average n was only 1.39 and the estimate of ρ_1 was actually negative. Poultry data involving larger numbers of progeny should yield more reliable estimates of the parameters and more precise estimates of the variances of the three estimated β's. Data are available consisting of five generations (progeny populations for 1952–56 inclusive) of White Leghorns previously described in detail by Yamada, Bohren and Crittenden [1957]. The trait considered is percent production to January 1, transformed to angles. The average n per dam over the five years is 7.1,

TABLE 1

Estimates of Pertinent Population Parameters and Number of Parents and Progeny Observed in Each of Five Years.

Year	$\hat{\rho}_1^*$	$\hat{\beta}_1$	$\hat{T}^*$	$\hat{\rho}^*$	No. sires	No. dams	Average n
1952	.064	.128	.059	.0557	10	106	5.7
1953	.026	.052	.029	.0282	10	92	9.5
1954	.0572	.096	.051	.0484	11	78	7.3
1955	.0688	.133	.055	.0521	18	108	7.9
1956	.0404	.089	.034	.0327	20	132	5.4
Ave.	.0513	.100	.046	.0434	—	—	7.1

varying from 5.3 in 1956 to 9.5 in 1953 (Table 1). Since the sires were mated for the season to a random group of dams the regressions are to be estimated on an intrasire basis. Consequently interest will center on estimating ρ_1^*, ρ^* and T^*. Analysis of variance in the hierarchal model was used to derive estimates of the variance components for sires (σ_s^2), dams (σ_d^2) and progeny within dams [$\sigma_0^2 = \sigma^2(1 - \rho)$]. From these components were derived estimates of ρ_1 (the full-sib correlation in this example) and preliminary estimates of h^2, for the purpose of estimating T^*, which in turn, is needed in selecting the appropriate weighting factors. These estimates were obtained from the variance components as

$$\hat{\rho}_1^* = \hat{\sigma}_d^2/(\hat{\sigma}_d^2 + \hat{\sigma}_0^2),$$

and

$$\hat{h}_1^2 = 4\hat{\sigma}_d^2/(\hat{\sigma}_s^2 + \hat{\sigma}_d^2 + \hat{\sigma}_0^2).$$

Each annual value of $\hat{h}_1^2$ so derived was considered as a preliminary estimate of 2β so $\hat{\beta}_1 = \hat{h}_1^2/2$. The estimates $\hat{\beta}_1$ and $\hat{\rho}_1^*$ were then used to obtain the estimates $\hat{T}^*$ and $\hat{\rho}^*$. The results are shown in Table 1. All values are relatively consistent from year to year and none of the values for $\hat{\rho}^*$ are outside the expected range based on the theoretical additive genetic model.

Since the value of T^* is unknown, the estimated value $\hat{T}^*$ derived from the data for each year was used in the formulae pertaining to the variances of the estimates. To approximate the minimum variance under the weighting technique (method 3) τ was set equal to $\hat{T}^*$. For the repeated parent technique (method 2), τ was assumed to be zero.

Estimates of β were obtained within each sire group in each year by each of the three estimation techniques. The variances of regressions in each sire group were estimated by use of the appropriate formula (i.e. 15, 17, or 18). The individual sire estimates of β were pooled over sires in years, using the reciprocals of the estimated variances as weights for the corresponding regression coefficients, to obtain a point estimate of β for each year. The estimated variance of the single point estimate of β for the year is the harmonic mean of the variances for each sire in the year divided by the number of sires in the year. Thus

$$\hat{\beta} = \sum_i (\hat{\beta}_i/\hat{\sigma}^2_{\beta_i}) / \sum_i (1/\hat{\sigma}^2_{\beta_i}),$$

and

$$\hat{\sigma}^2_{\beta} = 1/\sum_i (1/\hat{\sigma}^2_{\beta_i}).$$

To estimate heritability (h^2), the estimates of β were doubled. The estimates of h^2 and the estimated standard errors for each of the three methods of derivation are presented for each of the five years in Table 2. It is clear that in each year the estimated standard errors of the regression estimates based on progeny means (method 1) are the largest.

TABLE 2

HERITABILITY ESTIMATES AND STANDARD ERRORS, BASED ON REGRESSION COEFFICIENTS ESTIMATED BY THREE DIFFERENT METHODS.

	Method		
Year	1	2	3
1952	.156 ± .242	.317 ± .204	.281 ± .202
1953	−.006 ± .171	−.073 ± .166	−.058 ± .166
1954	.333 ± .304	.294 ± .297	.304 ± .296
1955	.110 ± .172	.098 ± .169	.102 ± .169
1956	.309 ± .216	.340 ± .198	.342 ± .198

There is little difference in the efficiencies of methods (2) and (3) in these data. Unless the data to be considered have larger values of ρ or ρ^* than observed in the present set of data, there appears to be little advantage in using the weighted technique (method 3) in preference to the repeated parent technique (method 2).

REFERENCES

1. Kempthorne, O. and Tandon, O. B. [1953]. The estimation of heritability by regression of offspring on parent. *Biometrics 9*, 90–100.
2. Yamada, Yukio, Bohren, B. B. and Crittenden, L. B. [1957]. Genetic analysis of a White Leghorn closed flock apparently plateaued for egg production. *Poultry Sci. 37*, 565–80.

QUERIES AND NOTES

D. J. Finney, *Editor*

164 NOTE: On a Formula for the Estimation of the Optimum Dressing of a Fertilizer

F. Pimentel Gomes,
University of São Paulo, São Paulo, Brasil.

If Mitscherlich's law

$$y = A[1 - 10^{-c(x+b)}]$$

adequately represents the yield y of a crop to which x units of a nutrient have been applied, the optimum dressing x^* (Pimentel Gomes [1953]) is given by the expression

$$x^* = (1/c) \log \frac{Awc}{t \log e} - b. \qquad (1)$$

Let x_u be a standard dressing of a nutrient and u the response to it; formula (1) can be written

$$x^* = (1/c) \log \frac{cx_u}{(1 - 10^{-cx}u) \log e} + (1/c) \log \frac{wu}{tx_u},$$

where, as before, w is the unit price of the crop yield, and t the unit price of the nutrient. This formula can be simplified through the expansion of its first term, as we shall show.

Let

$$z = cx_u/\log e,$$

$$Y = (1/c) \log \frac{cx_u}{(1 - 10^{-cx}u) \log e} = (1/c) \log \frac{z}{1 - e^{-z}}.$$

We have

$$\frac{dY}{dz} = (1/c)\left[\frac{1}{z} - \frac{1}{e^2 - 1}\right] \log e.$$

But it is known (Cramer [1946] p. 123) that

$$\frac{z}{e^z - 1} = 1 - \frac{z}{2} + B_2 \frac{z^2}{2!} + B_4 \frac{z^4}{4!} + \cdots,$$

where B_2 , B_4 , $\cdots$ are Bernoulli numbers. This power series converges for $\mid z \mid < 2\pi$, and therefore can be integrated with this restriction.

We obtain:

$$Y = (1/c) \log e \left(\frac{z}{2} - B_2 \frac{z^2}{2!2} + B_4 \frac{z^4}{4!4} - \cdots \right),$$

whence,

$$Y = x_u[(1/2) - (1/24)(cx_u/\log e) + (1/2880)(cx_u/\log e)^3 - \cdots].$$

The last series is alternating. If we keep only the first term, the error committed is less than $(x_u/24 \log e)\ cx_u$, that is, less than $cx_u^2/10$.

If two terms are kept, the error of the approximation is less than

$$(x_u/2880)(cx_u/\log e)^3.$$

Evidently, in applications, the first term gives a rather good approximation suitable for most cases, and this first approximation is independent of the parameters of the curve. To show how good this approximation is we give in Table 1 exact and approximate values of Y for the standard dressing (x_u) of 60 kg/hectare, which is suitable for most cases in practice.

TABLE 1

Values of c	Values of Y	(kg/hectare)
(hectares/kg)	Exact	Approximate
0.0020	29.4	30.0
0.0050	28.3	30.0
0·0090	26.9	30.0
0.0100	26.6	30.0

In most cases the approximate value

$$x^* = (1/2)x_u + (1/c) \log (wu/tx_u), \tag{2}$$

exceeds the true value, formula (1), by less than 5 percent.

From (2) we conclude also that, when the additional income (wu) produced by the nutrient applied is equal to the additional cost of fertilization (tx_u), the optimum dressing is still approximately $(\frac{1}{2})x_u$.

Let us suppose, now, that x^* will be estimated with the aid of a value of c obtained from a large group of previous experiments. Then we may take c as constant and obtained from either formula (1) or (2)

$$dx^* = (\log e/c)\,\frac{du}{u}\,;$$

hence an estimate of the variance of x^* is

$$V(x^*) = (\log e/c)^2\,\frac{2s^2}{ru^2}$$

where the response u is supposed to have been estimated by the difference between two means of r replications each. It is interesting to note that this estimate of $V(x^*)$ is independent of the unit prices w and t.

On the other hand, if u and c are estimated with data of the same experiment, then it is easier to use formula (2), which gives

$$dx^* = (\log e/\hat{c})\,\frac{du}{u} - \frac{d\hat{c}}{\hat{c}^2}\log\frac{wu}{tx_u}\,;$$

hence

$$V(x^*) \qquad (\log e/u\hat{c})^2\,V(u) + (1/\hat{c}^4)\left(\log\frac{wu}{tx_u}\right)^2 V(\hat{c})$$
$$- 2(\log e/\hat{c}^3)\left(\log\frac{wu}{tx_u}\right)\mathrm{Cov}\,(u, \hat{c}).$$

REFERENCES

Pimentel Gomes, F. [1953]. The use of Mitscherlich's regression law in the analysis of experiments with fertilizers. *Biometrics 9*, 498–516.

Cramèr, Harald [1946]. *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

BOOK REVIEWS

J. G. Skellam, *Editor*

Members are Invited to Suggest Books for Listing of Review to the Editor

12 EKAMBARAM, S. K. **The Statistical Basis of Quality Control Charts.** A Manual for Business and Factory Managers. Bombay and London: Asia, Publishing House, 1960. Pp x + 96. 16 Tables and 14 Diagrams. Rs. 6.50, 16s.6d.

L. R. Shenton, *University of Manchester, Manchester, England.*

This short introductory account of control theory and practice is intended for managerial and technical personnel, and also for initial courses at university level. After an account of frequency distributions leading to the normal law, the control chart for quality is described, and then the control of fraction defective, concluding with remarks on acceptance sampling. In describing the underlying statistical ideas Professor Ekambaram is unusually lucid and achieves considerable success, although at times the terminology is unconventional.

It seems unnecessary to have evaluated separately and in detail the first four moments of the Normal, Binomial and Poisson distributions; in any case little use is made of μ_3 and μ_4. On the other hand there is scant reference to the normal probability integral, and few values are quoted. The use of ${}_nC_r$ for the usual binomial coefficient $\binom{n}{r}$ may be confused with n times C_r , and the printing of $\sum x/N$ as $1/N \sum x$ seems unfortunate. The formula for a normal probability density is incorrectly printed on p.21. The printing is sometimes tenuous and certainly lacks uniformity.

13 LIEBERMAN, GERALD J. AND DONALD B. OWEN. **Tables of the Hypergeometric Probability Distribution.** Stanford: Stanford University Press, 1961. pp. vi + 726, Part I: 6 Tables; Part II: Tables; Appendix. $15.00.

R. A. Bradley, *The Florida State University, Tallahassee, Florida, U.S.A.*

This book of tables gives a comprehensive tabulation of the hypergeometric probability distribution together with discussions of typical applications. In the notation of the book
$p(x) = p(N, n, k, x) = k!n!(N - k)!(N - n)!/(k - x)!(n - x)!x!N!(N - k - n + x)!$ where $\max [0, n + k - N] \leq x \leq \min [n, k]$. $P(x)$ is used as the cumulative distribution. Symmetries are noted that permit reduction of the volume of the tables. Part II of the book and the Appendix consist of tables. The main section of tables, pp. 33–627, is a tabulation of $p(x)$ and $P(x)$ for $N = 2$, $n = 1$ through $N = 100$, $n = 50$. The second section of tables, pp. 628–705, shows values of the same quantities for $N = 1000$, $n = 500$. The third section of tables, pp. 706-713, gives $p(x)$

and $P(x)$ for $N = 100$, $n = 50$ through $N = 2000$, $n = 1000$ with $k = n - 1, n$; $n = N/2$. The Appendix consists of values of log $N!$ for $N = 1, \cdots, 2000$. The computations were effected on an IBM 704 computer and sufficient checks were used to insure the six-place accuracy given for all values of $p(x)$ and $P(x)$.

Part I of the book deals with comments and definitions relative to production of the tables, applications of the tables, and approximations to the hypergeometric probability distribution. The discussion of applications is clear, concise and useful. The examples given are Applications to a Sequential Procedure, Applications to Tests of the Equality of Two Proportions, Applications to the Distribution of the Number of Exceedances, Applications to the Binomial Distribution (Bayesian Prediction), and Applications to Sampling Inspection. A bibliography containing sixty-six references will be useful to users of these tables.

Tables of the Hypergeometric Probability Distribution will become an important reference volume and one recommended to all working in areas of relevance.

PILLAI, K. C. S. **Statistical Tables for Tests of Multivariate Hypotheses.**
14 Manila: Statistical Center, University of the Philippines. 1960. pp viii + 46.
M. J. R. HEALY, *Rothamsted Experimental Station, Harpenden, England.*

Most of the significance tests in common use can be regarded as special cases of the F-test for the equality of two estimated variances. The corresponding multivariate test would be one for the equality of two variance-covariance matrices, and the natural requirement that the result of the test should not depend on the scales of measurement implies that the test criterion must be some function of the latent roots (eigenvalues) of the two matrices. Denoting the matrices by $\mathbf{S}_1$ and $\mathbf{S}_2$, all tests so far proposed have in fact been based on the latent roots of the "ratio" $\mathbf{S}_1\mathbf{S}_2^{-1}$ or of $\mathbf{S}_1(\mathbf{S}_1 + \mathbf{S}_2)^{1}$. Wilks' Λ criterion is equal to the product of the roots, i.e. the determinant of the matrix, while other authors have suggested using the sum of the roots or the largest or smallest root.

Pillai's tables are for use with the last two criteria. Table 1 gives the upper 5% and 1% points of the largest latent root for s, the number of measurements, equal to 2(1)6, while Tables 2 and 3 give the same percentage points for the sum of the roots of $\mathbf{S}_1(\mathbf{S}_1 + \mathbf{S}_2)^{-1}$ and of $\mathbf{S}_1\mathbf{S}_2^{-1}$ respectively for $s = 2(1)8$. The last criterion is equivalent to Hotelling's T_0^2. Significance levels of the largest latent root have been tabulated very fully by F. G. Foster and D. H. Rees for $s = 2, 3$ and 4. Otherwise, the tables appear to break new ground. A short introduction includes several examples of multivariate problems to which the suggested tests can be applied.

The user of the tables is not given all the help he might expect. The parameters are not as convenient as the degrees of freedom in the Foster-Rees tables, and the parameter values make interpolation awkward. No guidance on interpolation is given in the introduction, and there is no indication of the accuracy of the tabular entries although these are based on approximations to the true distribution functions. In tables 2 and 3, the 5% and 1% values for a given s do not appear at a single opening of the book.

In spite of these shortcomings, the tables ought to be widely used, if only to provide experience on which to base further work. Several questions cannot at present be answered. In particular, no guidance can be given in choosing between the three alternative tests. Another open question concerns the robustness of the tests to non-normality; analogy with the ordinary F-test suggests that they may be unduly sensitive in this respect.

YATES, F. **Sampling Methods for Censuses and Surveys. 3rd Edition.**
15 London: Charles Griffin and Co. Ltd. 1960. Pp xvi + 440.54s.

M. D. Mountford, *The Nature Conservancy, London, England.*

This book is still being read after ten years since its first edition and has thus, according to the modern definition, achieved the status of a classic.

This, the third edition, is a reprint of the second edition with an extra chapter on the use of electronic computers in the analysis of censuses and surveys. The value of high-speed computers in large-scale survey work is now unquestioned; Dr. Yates' exposition of their workings and the timeless character of the earlier chapters brings this book right up to date.

The new chapter begins with a concise description of the main features of a computer and of the principles of constructing a programme of machine orders. He then presents a general programme for the analysis of survey results. This same general programme, with slight modifications, will also serve to instruct the computer to analyse the results of many different types of sampling schemes, including the simple random, stratified, ratio, regression and multiphase methods. The exposition is clear and unvarnished, though the reader may be dazzled by the simplicity of the unified treatment of the different sampling methods.

As a standard manual for the practical survey worker this book, to my knowledge, still has no equal.

ABSTRACTS

The following are abstracts of papers presented at meetings of the British Region on February 28 and April 18, 1961.

753 P. D. OLDHAM (M. R. C. Pneumoconiosis Research Unit, Llandough Hospital, Penarth, Glamorgan). **The Distribution of Arterial Pressure in the General Population.**

Blood pressure measurements of the general population form smooth distribution curves whose means and standard deviations vary with age and differ between the sexes. No other common factors influencing arterial pressure to a major extent have been discovered, nor does it appear, from second surveys of the same samples, that the distribution of change of pressure will materially depend on simple, common factors. The interpretation of these distributions raises the problem, occurring in all fields of medicine, of distinguishing normality from abnormality. The tendency is for unsatisfactory and arbitrary rules to be adopted for this purpose, rules which ignore the evident fact that abnormality cannot be diagnosed from the result of a single test, since the innumerable factors influencing function must be mutually correlated.

754 L. R. TAYLOR (Rothamsted Experimental Station, Harpenden, Herts.). **A Power Law Transformation for Aggregated Populations.**

The individuals of any species affect each other in many ways. The total effect of this attraction or repulsion appears, in the spacial distribution of the population, as a departure from the statistically simple ideal of randomness. The variance is affected and some powerful analytical technique are inapplicable. In set experiments, where the mean varies only 100 or 200% between treatments this can be overcome by a transformation devised for the occasion and not necessarily very effective at other population densities. To be effective in field work, where means may cover 6 or more log cycles, a transformation must have a sound basis.

Such a system of transformations has been found, empirically, which fits all data so far available. It derives from the hypothesis that variance is proportional to a fractional power of the mean ($s^2 = am^b$). Considerable evidence supports this; only 1 out of over 30 populations examined shows appreciable deviation. The index b appears to be much more specifically stable than the factor a which varies with sampling method, population trend etc., (which may be very local e.g. increase by reproduction). It is suggested that b is a specific Index of Aggregative Behaviour, present in all individuals, possibly influenced by environment but independent of population density and trends.

The transformation function is $\phi(m) = Q \int m^{-b/2}\, dm$ where $\phi(x)$ is the transformation for individual counts, Q is a constant and m and b are derived from the power law for variance. a disappears in transformation which therefore remains the same for the same species with different sampling practices in the material so far examined. (For further details see *Nature* (1961), *189*, 732–5.)

755 G. HARRINGTON (A. R. C. Statistics Group, School of Agriculture, Cambridge). **Studies of Visual Judgments of Quality in Bacon.**

The series of experiments to be described studied the manner in which experienced and naive judges handled their rating scale when visually assessing relatively simple characteristics of bacon. The sorts of attributes involved were the "proportion of lean to fat" on a cut surface revealed when a bacon side is cut into two halves, and the component features of this. Most experiments were carried out using photographs, in some cases mailed to the judges, although a few involved actual bacon sides. Balanced incomplete block arrangements were used so that the average quality of the various batches judged could be varied systematically. The analysis was concerned with the relative importance of variations in scores introduced by alterations of "judging standard" (position on rating scale) from batch to batch, judge differences etc., and the interrelations between various scores and measurements which may have inffuenced them.

756 J. M. TANNER and M. J. R. HEALY (Institute of Child Health, Gt. Ormond St., London, W. C. 1. and Rothamsted Experimental Station, Harpenden, Herts.). **Assessment of Maturity from X-rays of the Wrist and Hand.**

The bones of the wrist and hand pass through a number of distinguishable stages of growth during the period between birth and adulthood, and their overall state of development may be taken as a measure of the individual's level of physical development. This is often done by way of an assessment of "skeletal age". A scoring system for deriving skeletal age from an X-ray will be described.

The following are abstracts of papers presented at the meetings of E.N.A.R. held at Cornell University, Ithaca, N. Y., on April 20, 21, 22, 1961.

757 W. H. BEYER and R. E. BARGMANN (Virginia Polytechnic Institute, Blacksburg, Virginia). **Symmetrical Complementation Design.**

This design is intended for those experimental situations where the total amount of three treatments is a constant. The levels of the treatments are referred to a common unit of measurement, and are equally spaced. Certain cell entries are omitted to insure complete exchangeability of the three treatments. The usual additive model is assumed. This design differs from the usual types of design, in that the number of estimable contrasts is limited. Estimable functions in one treatment only and in two treatments are presented for the general case of p levels. Several methods are employed in order to obtain estimates of the treatment effects under various constraints. These estimates are rather meaningless quantities, as it is only when they are combined in estimable functions that unique results are obtained. Sums of squares and test statistics are presented for the various estimable hypotheses formulated. This paper shows that the "hypothesis of substitution" is one of the most important to consider. If accepted this says that applying one treatment at a low level and another at a high level does not produce results which are different when the two treatments are interchanged. Indication of how one might consider response functions for single treatments is given. The analysis is also extended to an analysis of covariance and then further to a multivariate analysis. Recommendations for interpretation and statement of limitations are made in detail.

758 A. E. BRANDT (Statistical Section, Agricultural Experiment Stations, University of Florida, Gainesville, Fla.). **The Analysis and Interpretation of Half-Replicate Experiments.**

An IBM 650 program for analyzing the data from a factorial experiment involving not more than 8 factors or independent variables or from an experiment which can be arranged in factorial form is presented. Of the 8 factors, 6 must have less than 10 classes and the remaining 2 may have 10 or more.

The 650 may be used to design a half-replicate experiment, that is, to designate the treatment combinations to be used, and to analyze the results. The output consists, in the case of a $(2)^n$ half-replicate experiment, of $(2)^n - 1$ cards. Of these, $2^n - 2$ contain the information concerning variances. These cards occur in pairs, one member of each being called an alias, on the basis of variances.

The data from a $(2)^2(4)^2$ experiment presented by W. H. Horton, Westinghouse Electric Corporation, to a seminar July 14, 1960 were analyzed by this program and the results submitted. A question is raised as to the interpretation of results. The effect of temperature level proved to be highly significant but one wonders if its alias, in this case a high order interaction involving the other 2-level factor and 2 levels of each of the 4-level factors, is to be ignored.

The results of a $(2)^3(4)^2$ field experiment over 4 years were analyzed and the results presented with both identifications given for each separate sum of squares. The question is again raised as to which member of a pair is to be accepted.

759 BYRON WM. BROWN, JR. (School of Public Health, University of Minnesota, Minneapolis, Minnesota). **Some Characteristics of the Spearman-Karber Estimator in Bioassay.**

Let the dose-response function in quantal assay be a distribution function with mean μ. An experiment for estimation of μ involves n subjects tested at each of the dose levels $x_i = x_0 + id$, $i = 0, \pm 1, \pm 2, \cdots$, where n, d and x_0 are fixed. The Spearman estimator is defined for this infinite experiment and shown to have finite mean and variance. Maxima are obtained for the bias, taken over all placements, x_0, of the dose mesh, for (i) all distribution functions and (ii) all unimodal distribution functions with specified maximum slope. These maxima are compared with the sequence of bounds obtained using Euler-MacLaurin formulae. The mean square error of the Spearman estimator is given for x_0 randomly chosen over $(0, d)$. The minimum mean square error of the estimator, for random choice of x_0 and fixed $n' = n/d$, occurs when $n = 1$ and $d = 1/n'$. As n' becomes infinite the estimator is consistent. The asymptotic variance of the estimator is defined and used to define asymptotic efficiency relative to the information. The only symmetric dose-response function, with μ as a translation parameter, for which the Spearman estimator has full asymptotic efficiency is the logistic distribution. There are distributions (with first moments) for which the Spearman estimator has asymptotic efficiency arbitrarily close to zero. High efficiencies are computed for the parametric models commonly used in bioassay. When the scale parameter is unknown the asymptotic efficiency of the Spearman estimator is at least that for the case of scale parameter known.

760 RICHARD G. CORNELL (The Florida State University, Tallahassee, Fla.). **Tables of Sample Sizes and Applications for Estimating some Monotonic Functions of the Ratio of Two Independent Poisson Variates.**

The calculation of the efficiency of a vaccine or of an air sampling device is an example when it is necessary to estimate monotonic functions of the ratio of two independent Poisson variates. Confidence limits on these efficiencies can be obtained by using the approach presented by Bross in "A Confidence Interval for a Percentage Increase," *Biometrics* (1954). It is also possible to use this approach to complete the sample size necessary to attain a confidence interval of predetermined length for any true efficiency. Sample sizes computed in this manner are tabled and applications are illustrated in this paper.

761 JAMES E. GRIZZLE (Dept. of Biostatistics, Univ. of North Carolina, Chapel Hill, N. C.). **Asymptotic Power of Tests of Linear Hypotheses Using the Probit or Logit Transformation.**

The statistic for testing the fit of a model, or the statistic for testing a linear hypothesis under the model, when using probits or logits, has a central χ^2-distribution for large samples if the null hypothesis is true. If it is not true, the test statistic has, asymptotically, a non-central χ^2-distribution with a non-centrality parameter that depends on the alternative hypothesis, the model and the transformation. Non-centrality parameters associated with tests of the two types of hypotheses are derived, and some cases of interest in bioassay when the response is quantal are examined.

762 JOHN GURLAND (Mathematics Research Center, U. S. Army, Univ. of Wisconsin, Madison, Wisc.). **Some Bioassay Techniques for the Determination of Minute Residues.**

A fortification procedure, whereby known amounts of the Standard material are added to weak test preparations, is suggested for biological assays involving housefly mortality and employing the "film method" and "topical method" respectively. According to the former method a thin film of toxicant is distributed on the walls of the jar in which the insects are exposed. Increasing doses are administered by increasing the volume of fortified extract employed in distributing the film of toxicant on the walls of the jar. According to the latter method, a constant volume (one microleter) is applied to the mesonotum of each fly exposed. This requires a separate fortification for each dose. By maintaining a constant ratio of toxicant added to toxicant present it is possible to apply the tests of linearity and parallelism and also to obtain an estimate of relative potency. Practical considerations in preparing solutions of desired strength (although the potency is unknown and must be estimated), are suggested, whereby a trial value of the relative potency is employed in obtaining the final estimate.

763 DEWEY L. HARRIS (Iowa State University, Ames, Iowa). **A Monte Carlo Study of the Influence of Errors of Parameter Extimation Upon Index Selection.**

The theory of genetic selection indexes is such that, with knowledge of certain genetic and phenotypic parameters, the index which will yield maximum genetic improvement may be chosen. However, in practice, these parameters are not known exactly and estimates are used in index construction. The inaccuracies of estimation result in indexes which will yield progress somewhat less than the maximum attainable progress. The errors of parameter estimation also lead to inaccuracies of estimating the progress from selection on a particular calculated index.

Parameter estimation from analyses of variance and covariance among traits of individuals classified into paternal half-sib groups was considered. The magnitude of the mean decrease in progress, the tendency to over- or under-estimate progress, and the accuracy of estimating progress were evaluated by "Monte Carlo" sampling procedures and by the development of approximate equations for various combinations of the true genetic and phenotypic parameters and amounts of data used for estimation. The results for these situations indicate that data involving at least 1000 individuals are necessary for construction of a reasonably effective index. However, the accuracy of estimating progress is not very accurate with this volume of data.

764 H. O. HARTLEY (Iowa State University, Ames, Iowa). **Analytic Studies of Sample Surveys.**

Analytic studies in sample surveys are concerned with estimating and comparing the mean-characteristics for certain sub-sections of the population called 'domains of studies'. The first part of the paper is concerned with formulas for the estimation of the totals and means of characteristics for all the units in a domain. Formulas of the variances of the estinates are also provided.

The second part of the paper raises the question of 'optimum design' for analytic studies and formulates this as a problem of minimizing the survey cost subject to tolerances for the variances of domain comparisons. Non-linear programming is applied to a simple special case in which domains are strata.

765 THEODOR HEIDHUES (Department of Animal Husbandry, Cornell University, Ithaca, New York). **Relative Accuracy of Selection Indices Based on Estimated Genotypic and Phenotypic Parameters.**

Empirical sampling techniques were used to investigate the effect of errors of parameter estimation on the accuracy of the selection index method. Under the assumptions of multivariate normal distribution of genotypic values and phenotypic observations and no genotype-environment interaction, samples of various sizes from two classes of underlying distributions were generated by an electronic computer. Genotypic values and phenotypic observations were computed such that expected values, variances and covariances of generated variables were equal to the respective underlying population parameters.

The first class of problems was concerned with indices which include phenotypic observations on an individual and its relatives in the same trait. The covariance matrix between genotypic values of relatives can be inferred from knowledge of the genetic mechanism and need not be estimated. The measure of the *relative accuracy* of an index based on estimated as compared to true parameters was taken to be the ratio of realized to expected correlation between genotypic value and its estimate by a particular index procedure. The decrease in accuracy due to use of estimated parameters depended upon the underlying ratio of genotypic to total variance and sample size. Evidence is strong that full utilization of genetic knowledge of the population structure increases the accuracy of an index procedure. Variables with low partial correlation with the genotypic value to be evaluated and high correlation with other variables in the index equation should be discarded from the index in certain cases. An index based on an estimated covariance matrix yielded almost identical accuracies when applied to the same sample or to different samples of the same population.

The second more general class of problems included indices based on phenotypic observations of the trait under selection and on a genetically correlated trait. The decrease in accuracy depended strongly on the ratio of genotypic to total variance of the selected trait. If estimates of elements of the phenotypic or genotypic covariance matrix are "unreasonable", i.e. if they exceed theoretically determined limits, they should be modified to increase the accuracy. The occasionally used practice of assuming the genotypic covariance between two traits to be zero cannot be generally recommended.

766 E. H. LEHMAN (Statistical and Computing Laboratory, Purdue University, Lafayette, Ind.). **The Peculiar Variance of the Estimator, of α, the Scale Parameter of the Weibull Distribution.**

Assume N units selected randomly from a population whose life span follows the Weibull density function, $F(t) = 1 - \exp[-(t^M/\alpha)]$, $t > 0$, M, (the shape parameter) known. Observe the failure times of the first few units and stop the test when R have failed and time T has elapsed. The maximum likelihood estimator $\hat{\alpha}$ of α derived from this test possesses some baffling properties. If R and T happen to be fixed and small, the variance of $\hat{\alpha}$ considered as a function of N, decreases at first, then increases for awhile, and later decreases monotonically. Thus it appears that for a certain interval of N, a small sample gives more information than a large one. The reason for this is that if r, the actual number of failures ($\geq R$, small) exceeds R by only a small integer, or if d, the actual test duration ($\geq T$, small), exceeds T by only a small period, the $\hat{\alpha}$ then employed is badly biased and variant. For small N, the probability of using these poor estimators grows for awhile before it shrinks, and hence the variance of $\hat{\alpha}$ as a whole follows this same surprising sinuous pattern.

767 DONALD C. MARTIN and S. K. KATTI (Florida State University, Tallahassee, Florida). **Fitting Certain Contagious Distributions to Some of the Available Data by Maximum Likelihood Method.**

The sample distributions obtained by Beal, [1940], *Ecology* 21, Bliss and Fisher, [1953], *Biometrics* 9, and McGuire *et al.*, [1957], *Biometrics*, 13 have been frequently employed to test the fit of many theoretical distributions. The distributions that have been found to have large enough regions of applicability are the Neyman Type A, Negative Binomial, Poisson Binomial, and the Inflated Poisson. All of these require numerical methods or tables to estimate the parameters by maximum likelihood. Specifically the problem of estimating parameters in the Neyman Type A and the Poisson Binomial is considerably involved. This has resulted in comparing of the newly formulated distributions with the inefficient fits, e.g. moment fits, of these distributions thereby confounding the superiority of the new distribution with the superiority of the method of estimation. The present authors have obtained maximum likelihood fits for most of the data that they feel are promising in studying the problem of curve fitting by using an electronic computer. They would like to report that the values of the chi squares used for testing the goodness of fit do not indicate any consistent good fit by one of these distributions.

768 M. R. SAMPFORD (A. R. C. Unit of Statistics, Aberdeen, Scotland). **A Problem in Cluster Sampling with Replacement.**

When the units of a population fall naturally into clusters of unequal size, selection of clusters with probability proportional to size, with replacement of selected clusters, provides easily calculated and unbiased estimators of the population mean and the variance of its estimate. Sampling usually proceeds until a pre-determined number n of clusters (not necessarily all different) has been chosen: under this system some economy of resources is achieved (since the mean of a cluster selected twice need be determined only once), at the cost of some increase in variance, possibly large when sampling is from a small population or stratum.

If resources to determine n cluster means are available, a sample containing n distinct clusters, and so providing a more precise estimate, might be preferred. If the usual method of sampling with replacement is continued until the $(n + 1)$th distinct unit is chosen for the first time, at the $(r + 1)$th drawing, the sample consisting of the first r cluster means (corresponding to the first n distinct clusters chosen) provides unbiased estimators when analyzed as though r (rather than n) had been pre-determined.

A small modification provides an unbaised estimator of the sampling variance when clusters are sub-sampled.

769 MARVIN A. SCHNEIDERMAN and PETER ARMITAGE (National Institutes of Health, and London School of Hygiene and Tropical Medicine, Bethesda, Md. and London, England). **A Family of Truncated Sequential Plans.**

For the normal deviate variance known, a family of truncated sequential plans (called "wedge" plans) with outer boundaries identical with those of Wald, Sobel-Wald, and Armitage have been developed. These plans have known, fixed Type I and Type II error and constitute a general class, of which the Wald (open) schemes are one extreme special case and the Armitage (restricted) schemes with a vertical middle boundary, the other extreme special case.

Plans have been developed for both the one-sided (two-decision), and for the two-sided (three-decision) case. Monte Carlo trials comparing "wedge" schemes with equivalent open schemes (Wald and Sobel-Wald) show somewhat increased average sample sizes for the wedge schemes in the vicinity H_0 , reduced sample sizes for values of the parameter, θ, between H_0 and H_1 ($\theta_1 > \theta_0$), and equivalent sample sizes for $\theta > H_1$. The variance of the ASN appears smaller for the wedge schemes at all values of the parameter. Tables of coordinates of the wedge for nine common combinations of α and β have been computed, and will be published.

770 R. J. TAYLOR and H. A. DAVID (Dept. of Statistics, Virginia Polytechnic Institute, Blacksburg, Va.). **Sequential Allocation of Patients in Clinical Trials.**

This paper describes a scheme for sequentially altering the proportion of patients assigned to the various treatments of a clinical trial according to the results obtained as the trial proceeds. Superior treatments are, by this means, allocated a higher proportion of patients than inferior ones. This alternation of proportion is performed by use of a weighting function, several alternative forms of which are given. A simulation study of the efficacy of the procedure with regard to its ability to select correctly the best treatment is described and the results presented. These results indicate that, with the use of appropriate weighting functions, this procedure is better able to select the best treatment than an equal allocation trial using the same number of patients. This comparison has been made on the basis of Sobel and Huyett's (1957) study of the equal allocation case. The study shows that the weighting functions which are most efficacious in correctly selecting the best treatment are the ones that tend to assign the largest proportion of patients to the best treatment.

A theoretical study of these statistics in special situations is also discussed.

771 L. H. WADELL (Department of Animal Husbandry, Cornell University, Ithaca, N. Y.). **Selection Bias in Intraclass Correlation Repeatability Estimates.**

Research workers in the field of quantitative genetics use functions of variance components as estimates of parameters needed in the design and application of selection programs. The estimates of these parameters generally have to come from selected data. This paper demonstrates the bias that is introduced into the intraclass correlation when computed from a one-way classification analysis with unequal subclass numbers where the unequal subclass numbers are caused by systematic truncation culling. Empirical sampling results are given to demonstrate this bias which varies with culling intensity and with the size of the true intraclass correlation. The bias introduced by increasing the culling intensity is greater for a low true repeatability than for a high true repeatability. A correction technique is given for this analysis which eliminates this bias. The accuracy of this method is supported by empirical sampling results.

772 R. M. ZAKI, B. B. BHATTACHARYYA and R. L. ANDERSON (North Carolina State College, Raleigh, N. C.). **On a Problem of Production Planning Over Time.**

This paper is concerned with the decision problem faced by a firm which produces a nonstorable commodity and has to spend large amounts of capital on a specialized factor that could be idled part of the time by fluctuations in production. In particular, it is assumed that the firm uses f resources, N_r units of the rth resource being available for production. These resources are to be allocated to one or more of T different time periods, n_{rt} units of the rth resource to the tth period. Each of these n_{rt} units can produce y_{rt} units of output at a price less direct variable costs of p_{rt}. The cost of the specialized factor for being available in any of the T time periods is approximated by a constant multiple of the maximum production in any one time period. The decision problem is to determine the allocation plan $\{n_{rt}\}$ which maximizes

$$Z = \sum_{r=1}^{f} \sum_{t=1}^{T} n_{rt}y_{rt}p_{rt} - c \operatorname*{Max}_{t} \left[\sum_{r=1}^{f} n_{rt}y_{rt} \right]$$

subject to the restrictions

$$\sum_{t=1}^{T} n_{rt} \leq N_r \quad \text{and} \quad n_{rt} \geq 0.$$

A linear programming formulation of the problem is given. A procedure is developed for finding the optimal solutions to the problem for each value of c in the interval $c \geq 0$. Explicit optimal solutions are given for a simplified model with $f = 1$.

The following are abstracts of papers presented to W.N.A.R. at the University of Washington, Seattle, Wash., June 14–17, 1961.

773 WALTER A. BECKER and LAWRENCE R. BERG (Washington State University, Pullman and Puyallup, Washington). **Factors Affecting the Sensitivity of Growth Experiments.**

A series of chick growth experiments utilizing various genotypes and diets were performed to investigate the influence of different factors upon the sensitivity of experiments. F statistics, M.S. among treatments/M.S. among individuals, within treatments in one-lay layouts, were used to measure sensitivity. All body weight data were transformed to logarithms.

In experiments that determined differences among diets, the within treatment variance increased as the animals grew older, reached a plateau, and then declined. The F value acted in a similar manner. The within treatment variance was greater for birds fed sub-optimal diets than for those fed optimal diets. When determining differences among strains, the highest F values occurred when birds were given the optimal diets. In nutritional research the "best" animals, in terms of producing the most sensitive experiment were those with highest nutritional requirements.

774 NEETI R. BOHIDAR (Iowa State University, Ames, Iowa). **Monte Carlo Investigations of the Effect of Linkage on Selection.**

A Monte Carlo investigation was undertaken to study the effect of linkage on the efficiency of selection. A program for the "Cyclone", the high speed digital computer located at Iowa State University, was written to accommodate any combination of the following facets: type of initial population, dominance, epistacy, linkage relations, selection intensity and some type of selection. The biological parameters involved in the actual numerical work were as follows: two types of population, repulsion and coupling; four types of dominance, no dominance, complete dominance over dominance, and mixed dominance; three types of character, character expressed by males, females and both sexes; three types of truncations, upper extreme, intermediate and lower extreme; two types of selection intensities, 20/40 and 5/40, and nine types of linkage relations, .5, .3, .1, .03, .015, .007, .003, t_f and t_m where t_f stands for tight linkage in female and t_m stands for tight linkage in male. Graphical method of representation was resorted to, to provide a clear picture of the situation. The results gave definite indications of the roles of these factors on the effects of selection and offered a comparative study of the effects of the combination of different facets of interest on the efficiency of selection.

775 H. D. BRUNK (University of Missouri, Columbia, Mo.). **A Statistic Related to Kolmogorov's.**

Let F_n denote the empiric distribution function of a random sample of size n from a population. In connection with distributions on a circle, Kuiper introduced the statistic $\max_x [F(x) - F_n(x)] - \min_x [F(x) - F_n(x)]$ [*Indag. Math.-Proc. Kon. Nederlandse Akad. Wet.*, Ser. A, 63 (1960) 38–47], for testing the hypothesis that the population has distribution function F. The statistic studied here is essentially Kuiper's: it bears the same relationship to Kuiper's as does Pyke's [*Ann. Math. Stat.*, Vol. 30 (1959), pp. 568–576] modification of Kolmogorov's to Kolmogorov's itself. The statistic occupies a position intermediate between Kolmogorov's [*Inst. Ital. Attuari, Giorn.*, 4 (1933) 1–11] and Sherman's [*Ann. Math. Stat.*, 21 (1950) 339–361], and the corresponding test appears more powerful than Kolmogorov's against certain alternatives (e.g. different scale parameter, for a symmetric distribution), and less powerful against others (e.g. different location parameter). Asymptotically the statistic coincides with that of Kuiper, who gives the asymptotic distribution (loc. cit.). A theorem of Sparre Andersen [*Skand. Aktuarietidskrift*, 36 (1953) 123–138] makes possible an essential simplification of the problem of determining the distribution for finite sample size. After this simplification, methods developed for Kolmogorov's statistic by Kolmogorov, Feller, Dempster and others can be used. Tables are in preparation.

776 JAMES L. LEITCH (Laboratory of Nuclear Medicine and Radiation Biology, School of Mecidine, University of California, Los Angeles, Cal.). **Radiation Effects and Their Statistical Evaluation: Introduction.**

With the ever-increasing interest in the biological effects of ionizing radiation, a review of the various facets in this field is considered as timely. It is necessary that all parameters, influenced by (1) radiation characteristics, (2) biological characteristics, (3) pre-irradiation conditions, (4) post-irradiation conditions, (5) treatment (protective) factors, and (6) criteria for evaluation of the biological effects, must be considered in any statistical appraisal. Initially these factors will be presented in outline form and discussed in more detail in subsequent papers.

777 JAMES L. LEITCH (Department of Nuclear Medicine and Radiation Biology, School of Medicine, University of California, Los Angeles, Cal.). **Biological Factors in Radiation Effects.**

A general review of literature data will be presented on the relationship between various biological factors and the radiation syndrome. Special emphasis will be placed on the treatment (protective) factors which may modify the basic syndrome. New experimental data involving X-ray effects on mice will be presented relative to the following parameters: (a) biological variability, (b) a possible seasonal effect, (c) cage effect and (d) interrelationship between dose rate and protection. An initial approach to the statistical evaluation of radiation experiments will be discussed.

778 FRANK J. MASSEY, Jr. and CARL E. HOPKINS (School of Public Health, University of California, Los Angeles, Cal.). **Tables of Exact Sampling Distribution of R2.**

The exact sampling distribution of the multiple correlation coefficient R^2 has been computed on the IBM 709 of Western Data Processing Center at UCLA and tabled for various magnitudes of the parameter, sample size, and number of variables. The density function and the cumulative distribution function are given in intervals of 0.01 in R^2, the sample coefficient.

These distributions should be useful in applications requiring significant tests and confidence intervals for sample R^2's where the hypothesis is non-zero, and for determining the power and sample size requirements of projected studies, such as epidemiologic surveys, in which the population R^2 is expected to be non-zero.

779 ORSELL M. MEREDITH (Laboratory of Nuclear Medicine and Radiation Biology, School of Medicine, University of California, Los Angeles, Cal.). **Comparison of Dose-Rate Effects on CF-1 Mouse: Mortality Between 250 KVP X-Rays and Cobalt-60 Gamma-Rays.**

A comparison of acute radiation mortality response with dose rates ranging from 2–170 r/min has been performed with CF_1 female mice. Analyses of response have been based upon methods of probit analysis elaborated by Finney. A rapid increase in the $LD_{50(30)}$ level was observed for either Co^{60} γ-rays or 250 KVP X-rays as the exposure dose rate was reduced below 20 r/min. On the other hand, little change in $LD_{50(30)}$ was observed with increase of exposure dose rate above 20 r/min. For either type of radiation source there was no significant deviation from parallelism when all of the probit dose response curves were considered. In addition, study has been made of the comparative applicability to Co^{60} γ-rays and 250 KVP X-rays results of various mathematical models which have been proposed for radiation mortality response.

780 STANLEY R. PERSON (Laboratory of Nuclear Medicine and Radiation Biology, School of Medicine, University of California, Los Angeles, Cal.). **Relationship Between Physical Characteristics of Radiation and its Biological Effect.**

Physical factors affecting the radiation sensitivity of whole animals will be discussed. Factors to be discussed center around changes in the radiation sensitivity brought about by use of radiation of different qualities. Data from the literature will be presented on the RBE of radiations that give rise to different rates of energy loss. A discussion of current dosimetry methods and factors affecting accurate dosimetry will be given.

781 A. D. WIGGINS (General Electric Company, Richland, Wash.). **Further Aspects of a Multicompartment Migration Model.**

The present paper represents an effort to extend the results of an earlier paper [Wiggins (1960). On a Multicompartment Migration Model With Chronic Feeding. *Biometrics 16:4*, 642–58] in several directions. First, the earlier model is generalized to include the possibility of an independent source or "feeding function" within each compartment. Second, a result of S. Bernstein [P. Lévy (1948). *Processus Stochastiques et Mouvement Brownien.* Gauthier-Villars, Paris. p. 64], namely the derivation of the one-dimensional diffusion equation of probability theory starting from a stochastic differential equation, is extended to K dimensions. The K-dimensional diffusion equation corresponding to the present migration model is then derived and an attempt is made to solve the equation in two dimensions.

Two numerical examples resulting from the application of the estimation procedure of the earlier *Biometrics* paper to experimental data are presented. The resulting graphs are plotted and compared with a plot of the experimental points.

CORRECTIONS

J. A. Nelder [1961]. The Fitting of A Generalization of the Logistic Curve. *Biometrics 17*, 89–110.

In the above paper the following reference on page 110 was omitted:

Skellam, J. G., Brian, M. V., and Proctor, J. R. [1959]. The simultaneous growth of interacting systems. *Acta Biotheoretica 13*, 131–144.

Also the algebraic expression in the heading of Table 1, (p. 91), should read

$$[1/(1 + e^{-\tau})].$$

R. C. Elston [1961]. On Additivity in the Analysis of Variance. *Biometrics 17*, 209-19.

The fourteenth line on page 215 should read: "null hypothesis, and, if condition (5) holds, provides an approximate."

THE BIOMETRIC SOCIETY

Brazilian Region

REPORT ON THE MEETING OF THE BRAZILIAN REGION

On March 29, 1961, the Brazilian Region of the Biometric Society held its annual meeting at the Department of Statistics of the Faculty of Hygiene and Public Health of the University of São Paulo, São Paulo, Brazil.

The first part of the meeting was devoted to the presentation of the following scientific papers:

"Cálculo da distância morfolótica em 4 grupos de Laelia Sp." by Rolando Vencovsky.

"Relação entre a Análise Tradicional de Experimentos Fatoriais de Adubação de 3^3 e a Superfície de Resposta Respectiva" by F. Pimentel Gomes.

"Análise de um experimento sôbre aplicação de micro-elementos em cafeeiros" by H. Vaz de Arruda.

"Correction for bias introduced by a transformation of variables" by Jerzy Neyman, as a visitor at the University of São Paulo.

The second part was devoted to the annual report for 1960 and the election of 1961 regional officers. The report and accounting demonstration for 1960 were submitted to the attendant members and approved.

In accordance with the results of the election, the names of new officers submitted for the approval of the Council of the Society are:

President—Adolpho M. Penha
Treasurer—Americo Groszman
Secretary—Elza Berquó
Council Members:
Pompeu Memória
José T. A. Gurgel
Frederico Pimentel Gomes
Ruy Aguiar da Silva Leme
Frederico G. Brieger
Geraldo Gracia Duarte.

British Region

At a meeting held on April 18th, 1961, the following papers were read and discussed:

G. Harrington: Studies of Visual Judgments of Quality in Bacon.

J. M. Tanner and M. J. R. Healy: Assessment of Maturity from X-rays of the Wrist and Hand.

W.N.A.R.

ANNUAL MEETING

The annual meeting of the WNAR Biometric Society was held at the University of Washington, Seattle, Washington, on June 14, 15, and 16, 1961, in conjunction with meetings of the Institute of Mathematical Statistics, the Section on Physical and Engineering Sciences of the American Statistical Association, the American Mathematical Society, and the Institute of Management Sciences; and a special "Symposium on Convexity" sponsored by the American Mathematical Society.

Program

Wednesday, June 14, 1961

7:30–10:00 p.m.—*Informal Inferential Procedures* (IMS, ASA-SPES and WNAR)
Chairman: A. M. Mood, C.E.I.R., Inc., Los Angeles, California

1. "The Future of Data Analysis." J. W. Tukey, Princeton University and Bell Telephone Laboratories, Murray Hill, New Jersey.
2. "Some Sequences of Fractional Replicates." C. Daniel, New York.
3. "Graphical Methods for Internal Comparisons in Multi-response Experiments." M. B. Wilk and R. Gnanadesikan, Bell Telephone Laboratories, Murray Hill, New Jersey.

Thursday, June 15, 1961

8:30–10:00 a.m.—*Stochastic Processes in Biology*
Chairman: A. T. Bharucha-Reid, University of Oregon

1. "Further Aspects of a Multi-Compartment Migration Model." A. D. Wiggins, General Electric Company.
2. "A Statistic Related to Kolmogorov's." H. D. Brunk, University of Missouri.

10:30 a.m.–12:30 p.m.—*Planning and Analysis of Experiments* (ASA-SPES and WNAR)

1. "The Consideration of Variance and Bias Errors in the Selection of a Response Surface Design." G. E. P. Box, University of Wisconsin, and N. R. Draper, Mathematics Research Center, U. S. Army.
2. "Orthogonal Main-Effect Plans." Sidney Addelman, Research Triangle Institute.
3. "Asymmetric Factorial Designs and the Direct Product." B. Kurkjian, Diamond Ordnance Fuze Laboratories, and M. Zelen, University of Maryland and National Bureau of Standards.

2:15–3:45 p.m.—*Radiation Effects and their Statistical Evaluation*
Chairman: James L. Leitch, Laboratory of Nuclear Medicine and Radiation Biology, U.C.L.A.

1. Introduction
2. "Relationship between Radiation Characteristics and Radiation Effects." S. R. Person, Laboratory of Nuclear Medicine and Radiation Biology, U.C.L.A.
3. "Biological Factors Involved in Radiation Effects." James L. Leitch.
4. "Comparison of Dose-Rate Effects on CF-1 Mouse Mortality Between 250 KVP X-Rays and Cobalt-60 Gamma Rays." Orsell M. Meredith, Laboratory of Nuclear Medicine and Radiation Biology, U.C.L.A.

Friday, June 16, 1961

8:30–10:30 a.m.—*Estimation* (IMS and WNAR)

1. "Combining Information in Incomplete Blocks." F. A. Graybill, Colorado State University.

2. "Estimation with Minimum Mean Square Error." H. O. Hartley, Iowa State University.
3. "Remarks on the Efficiency of Unbiased Estimation with Auxiliary Variates." W. H. Williams, Bell Telephone Laboratories, Murray Hill, New Jersey.

1:00–2:30 p.m.—*Contributed Papers*

Chairman: A. D. Wiggins, General Electric Company

1. "Monte Carlo Investigation of the Effect of Linkage on Selection." N. R. Bohidar, Utah State University.
2. "Tables of Exact Sampling Distribution of R^2." F. J. Massey, Jr. and Carl E. Hopkins, U.C.L.A.
3. "Various Levels of Riboflavin and the Sensitivity of Experiments on Growth." W. A. Becker and L. R. Berg, Washington State University.

CHANGES IN MEMBERSHIP

(January 15–July 15, 1961)

Changes of Address

Mr. Ross W. Adams, 1251 Hawthorne, Ames, Iowa, U.S.A.

Mr. B. L. Adkins, Statistics Department, University of New England, Armidale, N.S.W., Australia.

Miss Margaret F. Allen, School of Aviation Medicine, USAF, Brooks AFB, Texas, U.S.A.

Mr. H. A. J. Amand, 21 Place Cardinal Mercier, Rizensart, Belgium.

Mr. Donald W. Bailey, Cancer Research Institute, Univ. of California Medical Center, San Francisco 22, California, U. S. A.

Mr. B. O. Bartlett, Agricultural Research Council, Letcombe Regis, Wantage, Berkshire, England.

Dr. Glenn E. Bartsch, Department of Preventive Medicine, Western Reserve University, Cleveland 6, Ohio, U. S. A.

Mrs. Hannelore Beyer, Haertelstr. 16–18, Leipzig C 1, Germany.

Mr. Paul Blunk, 4616 Plantation Drive, Fair Oaks, California, U. S. A.

Mr. W. F. Bodmer, Department of Genetics, University of Cambridge, Cambridge, England.

Mr. Roger L. Bollenbacher, 860 Hiawatha Drive, Elkhart, Indiana, U. S. A.

M. Jacques Bredas, 24 rue Grand Bry, Montiguy-Le-Tilleul, Belgium.

Dr. Leroy S. Brenna, The Texas Company, 12th Floor Chrysler Bldg., New York, N. Y., U. S. A.

Dr. A. Brown, Department of Mathematics, Australian National University, Canberra City, A.C.T., Australia.

Dr. Robert V. Brown, Box 181, Edgewood, Maryland, U. S. A.

Dr. W. R. Buckland, The Exonomist Intelligence Unit, St. James, London, S.W. 1, England.

Mr. A. Burny, 77 avenue des Combattants, Gembloux, Belgium.

Mr. Lyle D. Calvin, Department of Statistics, Oregon State University, Corvallis, Oregon, U. S. A.

Dr. A. H. Carter, 1 Hackin Place, Fairfield, Hamilton, New Zealand.

Mr. Melvin W. Carter, Department of Mathematics and Statistics, Purdue University, Lafayette, Indiana, U. S. A.

Mr. David B. Christian, 33 Crestview Drive, Whitesboro, New York, U. S. A.

Mr. Frank B. Cramer, 17331 Tribune Street, Granada Hills, California, U. S. A.

Mr. Mare Dalebroux, c/o Instituto di Genetica, Universita di Pavia, Pavia, Italy.

Dr. James G. Dare, Department of Pharmacy, University of Queensland, Brisbane, Australia.

Dr. Richard J. Daum, 3900 Hamilton Street, Hyattsville, Maryland, U. S. A.

Miss M. E. Davis, Department of Agriculture, Box 1500, Wellington, New Zealand.

Mr. R. De Coene, 103 rue Edith Cavell, Bruxelles 18, Belgium.

M. Jean Dejardin, ORSTOM, 24 rue Bayard, Paris (VIII^{0}) France.

Mr. R. Delhaye, 36 avenue Jean Van Raelen, Bruxelles 16, Belgium.

Dr. Daniel B. De Lury, Department of Mathematics, University of Toronto, Toronto 5, Canada.

Mr. A. Deville, 2 rue Middelbourg, Boitsfort, Belgium.

Mr. H. M. Dicks, Department of Agriculture, J. S. Marais Bldg., Stellenbosch, South Africa.

M. Pol Dineur, Golsinnes, Bossiere par Masy, Belgium.

Miss Irene L. Doto, Communicable Disease Center, 2082 West 38th Street, Kansas City 3, Kansas, U. S. A.

Dr. D. B. Duncan, Department of Biostatistics, The John Hopkins University, Baltimore 5, Maryland, U. S. A.

Mr. Steve A. Eberhart, Department of Agronomy, Iowa State University, Ames, Iowa, U. S. A.

Dr. F. Ectors, 111 rue du Centre, Assesse, Belgium.

Mrs. Polly Feigl, Olof Skotkonungsgatan 66, Goteborg S. Sweden.

Dr. Heinz Fink, Morgengraben 14, Koeln-Stammheim, Germany.

Mr. Robert Fitzpatrick, 5229 21st Avenue, N.E., Seattle 5, Washington, U. S. A.

Dr. Henry R. Fortmann, Agricultural Experiment Station, Pennsylvania State University, University Park, Pennsylvania, U. S. A.

Mr. Robert A. Harte, Am. Soc. of Biological Chemists, 9650 Wisconsin Avenue, Washington 14, D. C., U. S. A.

Prof. Dr. Jo Hartung, Ruehlmannstr. 8, Hannover, Germany.

Dr. Don W. Hayne, Patuxent Wildlife Research Center, Laurel, Michigan, U. S. A.

Prof. Dr. J. Hemelrijk, Keizer Karelweg 83, Amstelveen (N.H.), Netherlands.

Mr. Jean Henry, 1 rue Defacqs, Bruxelles, Belgium.

Dr. Paul G. Homeyer, C-E-I-R, Inc., 11753 Wilshire Blvd., Los Angeles 25, California, U. S. A.

Dr. Carl E. Hopkins, School of Public Health, University of California, Los Angeles 24, California, U. S. A.

Mr. Paul V. Hurt, Deerfield, Wisconsin, U. S. A.

Dr. Peter Ihm, c/o Euratom, Casella postale 191, Como, Italy.

Mr. Arthur G. Itkin, 1870 Clayton Road, Abington, Pennsylvania, U. S. A.

Mr. Willaim G. H. Ives, Forest Biology Laboratory, Box 6300, Winnipeg, Manitoba, Canada.

Dr. Dubodh K. Jain, Botany Division, Indian Agricultural Research Institute, New Delhi 12, India.

Mr. Eugene A. Johnson, Industrial Engineering, University of Minnesota, Minneapolis, Minnesota, U. S. A.

Dr. med. Herbert Jordan, Haus Tusculum, Bad Elster, Germany.

Prof. Dr. Hans Kelleher, Ludwigstr. 18, Munchen, Germany.

Mr. Thomas R. Konsler, Mt. Hort. Crops Research Station, Route 2, Fletcher, North Carolina, U. S. A.

Dr. Paul Kuehne, Nachodstr. 19, Berlin W 30, Germany.

Mr. Thomas E. Kurtz, Mathematics Department, Dartmouth College, Hanover, New Hampshire, U. S. A.

Mrs. Katherine B. Ladd, 4 Woodside Drive, South Burlington, Vermont, U. S. A.

Dr. R. J. Ladd, Physiology Department, University of Queensland, Brisbane, Australia.

Dr. G. E. J. Lambelin, 231 Chaussee d'Alsemberg, Bruxelles 18, Belgium.

Dr. Lonnie L. Lasman, A. J. Wood Research Corporation, 42 South 15th Street, Philadelphia 2, Pennsylvania, U. S. A.

Mr. Jay P. Leary, Jr., 1253 S. Longwood Avenue, Los Angeles 19, California, U. S. A.

M. Jerome Lejeune, Institut de Progenese, 15 rue de l'Ecole Medecine, Paris (VI^0), France.

Mr. Robert Lichter, Ragis-Stat. Heidehof, Brockhoefe, Kr. Velzen, Germany.

Dr. D. Lindley, Department of Statistics, University College of Wales, Aberystwyth, Cards, Wales.

Mr. George F. Lunger, P. O. Box 583, Camden, New Jersey, U. S. A.

Marcel J. W. Luttgens, 7 rue Van Oost, Brussels 3, Belgium.

Prof. Dr. A. Maede, Grasse Steinstr. 81 Halle/Saale, Germany.

Mr. James E. Mangan, 7443 N. Claremont Avenue, Chicago 45, Illinois, U. S. A.

Mr. Stuart H. Mann, 511 W. University, Champaign, Illinois, U. S. A.

Mr. Robert Marechal, J. M., 29 rue Docqs, Gembloux, Belgium.

Dr. M. Maricz, 16 avenue des Abeilles, Ixelles, Belgium.

Mr. T. J. Marynen, 11 ring Laan, Berchem-Anvers, Belgium.

Mr. John W. Mayne, Operational Research Group, Defence Research Board, Ottawa, Canada.

Mr. Judson U. McGuire, European Parasite Laboratory, 20 bis rue Sadi Carnot, Nanterre (Seine) France.

Mr. Martin Menzi, Kreuzrain, Hedingen (ZH) Switzerland.

M. Philippe Merat, 20 rue de Louvre, Viroflay (S. and O.) France.

Mr. Donald L. Meyer, 1646 S. State Street, Syracuse 5, New York, U. S. A.

Dr. A. M. Mood, C-E-I-R, Inc., 11753 Wilshire Blvd., West Los Angeles, California, U. S. A.

Prof. Sigeiti Moriguti, Department of Statistics, Stanford University, Stanford, California, U. S. A.

Dr. M. B. Mueller, Plastics Division, Allied Chemical Corporation, Glenolden, Pennsylvania, U. S. A.

Dr. Hugo Muench, Jr., 100 Memorial Drive, Cambridge 42, Massachusetts, U. S. A.

Dr. Karl Heinz Muller, F. Schelling Str. 3, Jena, Germany.

Mr. August Carl Nelson, Jr., 1015 Green Street, Durham, North Carolina, U. S. A.

Dr. A. R. G. Owen, Department of Genetics, University of Cambridge, Cambridge, England.

Dr. Erich Panse, c/o Lochow-Pettkus CmbH, Bergen Krs. Celle, Germany.

Dr. Benjamin Pasamanick, Columbia Psychiatric Institute and Hospital, Columbus 10, Ohio, U. S. A.

Dr. Mary Ellen Patno, 2451 Yost Boulevard, Ann Arbor, Michigan, U. S. A.

Mr R. Pierlot, 5 avenue des Phalenes, Bruxelles 5, Belgium.

M. Jacques Poly, Service de genetique animale, 16 rue de l'Estrapede, Paris (V^0) France.

Mr. Joe Powell, 8245 Park Place Blvd., Apt. 6, Houston 17, Texas, U. S. A.

Mr. Wolf Prensky, Department of Biology, Brookhaven National Laboratory, Upton, L.I., New York, U. S. A.

Mr. Lester W. Preston, Jr., A. H. Robins Co., Inc., 1407 Cummings Drive, Richmond 20, Virginia, U. S. A.

Mr. Dieter Rasch, Freiligrathstr. 14, Rostock, Germany.

Dr. Arthur Ringoet, 1 avenue du Congo, Bruxelles 5, Belgium.

Mr. Erwin Roth, Jaminstr. 30, Erlangen, Germany.

Dr. O. K. Sagen, 210 Fifth Street, S.W., Washington, D. C., U. S. A.

Mr. Wilfred Salhuana, Universidad Agraria, Apartado 456, La Molina, Lima, Peru.

Dr. Hellmut Schmalz, Berliner Str. 2, Hohenthurm-Saalkreis, bei Haale/Saale, Germany.

Cand. Math. Berthold Schneider, Marburger Str. 18, Giessen, Germany.

Dr. Francesco Sella, Instituto di Genetica, Via S. Epifanio 14, Pavia, Italy.

Mr. William Seyffert, MPI f. Zuechtungsforsch, Post Bickendorf, Koeln-Vogelsang, Germany.

Dr. Robert R. Shrode, 111 E. State Street, Sycamore, Illinois, U. S. A.

Dr. Donald F. Starr, Route 1, Box 321 A, Grand Island, Nebraska, U. S. A.

Mr. Otto Steiner, St. Wenderlstr. 50, Braunschweig, Germany.

Mr. N. S. Stenhouse, Division of Math. Statistics, C.S.I.R.O., University of Adelaide, Adelaide, S. Australia.

Dr. Klaus J. Stern, Manhagener Allee 84, Schmalenbeck/Ahrensburg, (Holst.) Germany.

Miss Elizabeth Street, 231 West 13th Street, New York 11, N. Y., U. S. A.

Mr. R. C. Tomlinson, 74 Gayton Road, Harrow, Middlesex, England.

Dr. M. Torfs, 99Bd Lambermont, Bruxelles 3, Belgium.

Mr. Gerard Torreele, 5 Slachthuesstraat, Niouwpoort, Belgium.

Dr. Jean Vacher, 22 avenue Grammont, Tours (Indre-et-Loire) France.

M. Raymond Van Den Driessche, 42 rue du Friquet, Bruxelles 17, Belgium.

Mr. A. Van Parijs, 128 rue de la Loi, Bruxelles 4, Belgium.

Mr. Thierry Waffelaert, 5 rue J. B. Verlooy, Anvers, Belgium.

Prof. Dr. Erna Weber, Schenkestr. 8c, Berlin-Karlshorst, Germany.

Mr. Irving Weiss, The Mitre Corporation, Bedford, Massachusetts, U. S. A.

Mr. Robert White, Box. 241, Dugway, Utah, U. S. A.

Mr. Henry K. C. Woo, E-TAI Ltd., 95 Liberty Street, New York 6, N. Y., U. S. A.

Dr. Gunter Wricke, ueber Lingen/Ems, Klausheide, Germany.

New Members

At Large

Dr. Martin Eugene Dehousse, University of Ruanda-Urundi, B.P. 1550, Usumbura-Burundi, Africa.

Mr. Hong Suk Lee, Last Crop Section, Agricultural Experiment Station, Suwon, Korea.

Mr. Heliodoro Miranda M, Inter-American Institute of Agricultural Sciences, Turrialba, Costa Rica.

Ing. Luis A. Montoya-Armas, Instituto Interamericano de Ciencias Agricolas, Turrialba, Costa Rica.

Australia

Prof. John Henry Bennett, Department of Genetics, University of Adelaide, Adelaide, South Australia.

Dr. B. Diamantis, 2 Raleigh Street, Windsor S. 1, Victoria, Australia.

Mr. Alan E. Stark, Div. of Fisheries and Oceanography, C.S.I.R.O., P. O. Box 21, Cronulla, N.S.W., Australia.

British

Dr. P. F. D'Arcy, Pharmacology Department, Allen and Hanburys Ltd, Ware, Herts., England.
Mr. J. P. Evenson, Wellcome Research Laboratories, Langley Court, Beckenham, Kent, England.
Mr. P. Hallam, 67 Tomline Road, Felixstowe, Suffolk, England.
Mr. D. J. Harberd, Plant Breeding Station, Pentlandfield, Roslin, Midlothian, England.
Mr. P. Holgate, "High Canons", Well End, Barnet, Herts., England.
Mr. G. J. Knight, Wellcome Research Laboratories, Langley Court, Beckenham, Kent, England.
Mr. Ian McDonald, 6 Deeside Place, Aberdeen, Scotland.
Mr. F. M. O'Carroll, 12 Dundela Avenue, Sandycove, Co. Dublin, Ireland.
Prof. W. T. Williams, Botany Department, The University, Southampton, England.

Belgian

Mr. Joseph J. Gabriel, 54 avenue Dr Decroly, Uccle-Bruxelles 18, Belgium.
Mr. Georges Geortay, 26/60 Avenue Georges Truffaut, Liege, Belgium.
Dr. L. Goeminne, 98 Chaussee de Gand, Deinze, Belgium.
Dr. Paul Janssen, Department de Recherches des Laboratories Pharmaceutique, Turnhout, Belgium.
Dr. A. Jeurissen, Sanatorium, Buizingen, Belgium.
Mr. Francois R. Martin, 5 Rue Caroly, Ixelles, Bruxelles, Belgium.
Mr. Andre Pieteres, Brusselse steenweg 407, Gentbrugge, Belgium.
Mr. Francois Sterckx, Yangambi II, B.P. 1035, Stanleyville, Belgian Congo.
Dr. Robert Van Vaerenbergh, 6 avenue du Saleil, Knokke, Belgium.

ENAR

Dr. Helen Abbey, 615 N. Wolfe Street, Baltimore 5, Maryland, U. S. A.
Dr. Elliott T. Adams, P. O. Box 47, Upham's Corner Station, Boston 25, Massachusetts, U. S. A.
Dr. Daniel J. Baer, 2877 Valentine Avenue, New York 58, N. Y., U. S. A.
Dr. Harle V. Barrett, Department of Preventive Medicine, The Creighton University School of Medicine, Omaha 2, Nebraska, U. S. A.
Dr. A. F. Bartholomay, 12 Upland Road, Wellesley, Massachusetts, U. S. A.
Miss Virginia B. Berry, Department of Mathematics, University of British Columbia, Vancouver B.C., Canada.
Mr. Paul V. Blair, Populations Genetics Institute, Purdue University, Lafayette, Indiana, U. S. A.
Dr. John R. Braunstein, 2123 Luray Avenue, Cincinnati 6, Ohio, U. S. A.
Mr. Franklin W. Briese, 7500 Olivers Avenue South, Minneapolis 23, Minnesota, U. S. A.
Mr. Paul M. Cohen, Technical Operations Inc., Box 37, Fort Monroe, Virginia, U. S. A.
Mrs. Elizabeth F. Davis, Hazleton Laboratories Inc., Box 30, Biometrical Unit, Falls Church, Virginia, U. S. A.

Dr. Roscoe A. Dykman, Division of Behavorial Sciences, University of Arkansas, Little Rock, Arkansas, U. S. A.

Mr. John R. Flood, 94 Parkway Road, Bronxville, New York. U. S. A.

Dr. D. H. Fogel, 1380 Bedford Street, Stamford, Connecticut, U. S. A.

Dr. Seymour Geisser, Biometrics Branch, NIMH, Bethesda, Maryland, U. S. A.

Dr. William T. Ham. Jr., Box 877, Medical College of Virginia, Richmond, Virginia, U. S. A.

Mr. Herman B. Hamot, 835 Forbes Avenue, Perth Amboy, New Jersey, U. S. A.

Dr. Dewey L. Harris, Statistical Laboratory, Iowa State University, Ames, Iowa, U. S. A.

Dr. Edwin Hendler, 415 Brentwood Road, Havertown, Pennsylvania, U. S. A.

Dr. Homer C. Jamison, 1919 Seventh Ave., S., Birmingham 2, Alabama, U. S. A.

Mr. Denis J. Kelleher. Department of Animal Husbandry, Iowa State University, Ames, Iowa, U. S. A.

Dr. Samuel J. Kilpatrick, Statistical Laboratory, Iowa State University, Ames, Iowa, U. S. A.

Mr. Eugene Legler. Tennessee Game and Fish Commission. Cordell Hull Building, Nashville, Tennessee, U. S. A.

Dr. Guillermo Llanos-Bejarano. 608 N. Collington Avenue, Baltimore 5, Maryland, U. S. A.

Mrs. Ruth B. Loewenson, 4844 Xerxes Avenue, S., Minneapolis 10, Minnesota, U. S. A.

Dr. Josiah Macy. Jr., Department of Physiology, Albert Einstein College of Medicine, New York 61, N. Y., U. S. A.

Mr. Robert H. Miller, Dairy Herd Improvement, Agricultural Research Service, USDA, Washington 25, D. C., U. S. A.

Dr. Richard Moore. American National Red Cross, 18th and E., N.W., Washington, 6, D. C., U. S. A.

Dr. Donald F. Morrison, Biometrics Branch, National Institutes of Health, Bethesda 14, Maryland, U. S. A.

Mrs. Sue W. Nealis. 6657-24th Place. Riggs Manor, Hyattsville, Maryland, U. S. A.

Dr. Masatoshi Nei, Department of Genetics, North Carolina State College, Raleigh, North Carolina, U. S. A.

Mr. Gill Nestel, 1219 S. State Street, Ann Arbor, Michigan, U. S. A.

Mr. Marcelo M. Orense, Department of Experimental Statistics, North Carolina State College, Raleigh, North Carolina, U. S. A.

Mr. James G. Osborne, Forest Service, USDA, South Building, Washington, D. C., U. S. A.

Dr. Bernard S. Pasternack, New York University Medical Center, 550 First Avenue, New York 16, N. Y., U. S. A.

Dr. H. V. Pipberger, 7439 Little River Pike, Annandale, Virginia, U. S. A.

Mr. P. V. Rao, Department of Mathematics, University of Georgia, Athens, Georgia, U. S. A.

Mr. Searle B. Rees, 9 Strathmore Road, Brookline, Massachusetts, U. S. A.

Dr. Richard D. Remington, School of Public Health, University of Michigan, Ann Arbor, Michigan, U. S. A.

Mr. J. C. Richards, Jr., The Standard Oil Company, Midland Bldg., Cleveland 12, Ohio, U. S. A.

Mr. Richard H. Richardson, 156 Williams Hall, N. C. State College, Raleigh, North Carolina, U. S. A.

Mr. Donald C. Riley, American Statistical Association, 1757 K Street, N.W., Washington 6, D. C., U. S. A.

Dr. Ralph Rossen, L. E. Phillips Psychobiological Research Division, Mt. Sinai Hospital, Minneapolis 4, Minnesota, U. S. A.

Mr. Raymond E. Roth, Department of Mathematics, St. Bonaventure University, St. Bonaventure, New York, U. S. A.

Mr. Jagdish S. Rustagi, College of Medicine, University of Cincinnati, Cincinnati 19, Ohio, U. S. A.

Mr. Darshan Lal Sachdeva, 308 South Macomb, Tallahassee, Florida, U. S. A.

Mr. Henry E. Schaffer, Department of Genetics, N. C. State College, Raleigh. North Carolina, U. S. A.

Mr. Marvin A. Schneiderman, Biometry Branch, National Cancer Institute, Bethesda 14, Maryland, U. S. A.

Mr. Wilfred M. Schutz, Department of Genetics, N. C. State College, Raleigh, North Carolina, U. S. A.

Mr. Edward Selig, 130 Beach Avenue, Mamaroneck, New York, U. S. A.

Mr. John L. Seliskar, U. S. Forest Service, USDA, South Building, Washington 25, D. C., U. S. A.

Dr. C. W. Sheppard, Department of Physiology, University of Tennessee, Memphis 3, Tennessee, U. S. A.

Mr. Robert E. Sherman, 4143 Blaisdell Avenue S., Minneapolis 9, Minnesota, U. S. A.

Mr. Lawrence E. Sly, Jr., 202 West 9th Avenue, Tallahassee, Florida, U. S. A.

Mr. Herbert Stern, Jr., 737 Carol Marie Drive, Baton Rouge 6, Louisiana, U. S. A.

Dr. Claire M. Vernier, Department of Medicine and Surgery, VA Central Office, Washington, D. C., U. S. A.

Dr. Richard L. Willham, 3624 Ross Road, Ames, Iowa, U. S. A.

Mr. Ralph P. Winter, 5712-38th Avenue South, Minneapolis 17, Minnesota, U. S. A.

Mr. Donald F. Wilson, 6127 Westchester Drive, Washington 22, D. C., U. S. A.

Dr. Charles Wunder, Department of Physiology, State University of Iowa, Iowa City, Iowa, U. S. A.

French

M. Jean Louis Beaumont, 14 rue Petrarque, Paris 16e, France.

M. Marcel Brunard, 17 avenue Emile-Deschanol, Paris 7e, France.

M. Paul Damiani, Institut National de la Statistique, 29 quai Branly, Paris 7e, France.

Mme. Jacqueline Roquet, Docteur en Pharmacie, 22 avenue Victoria, Paris 1^0, France.

M. Luu-Mau-Thanh, 19 Boulevard Brune, Paris (XIV^0) France.

German

Dr. K. H. Barocka, Gartenstr. 6, Einbeck/Hann., Germany.

Prof. Dr. B. Baule, Nibelungenstr. 63, Graz/Oesterreich, Austria.

Dr. W. D. Froehlich, Rochusweg 12, Bonn, Germany.

Dr. H. G. Kmoch, Inst. f. Pflanzenbau, Katzenburgweg 5, Bonn/Rhein, Germany.

Dr. J. Krippl, Nordendstr. 14, Munchen 13, Germany.

Dr. R. Krussmann, Anthropol. Institut, University, Mainz, Germany.

Dr. Gunter Wricke, Post Nordhorn, Fa. v. Lochow-Petkus, Klausheide, Germany.

India

Mr. Prem Narain, Animal Genetics Division, Indian Veterinary Research Institute, Izatnagar, U.P., India.

Mr. G. Narasimharao, Asst. Physiologist, Sugarcane Research Station, Anakapelle, India.

Mr. J. S. Ramaratnam, 293 '7 Suifabad Lane, Khairabad, Hyderbad, India.

Japan

Mr. Masaki Horie, Haraikata-machi 9, Shinjyuku-ku, Tokyo, Japan.

Mr. Kiyoo Kimura, Department of Mathematics, Mie Prefectural University, 11 Ootani-cho, Tsu Mie, Japan.

Mr. Kozuo Kitamura, Saitama Prefectural Agricultural Experiment Station, Ageo City, Saitama, Prefecture, Japan.

Mr. Masahiko Sugimura, Kumamoto Women's University, Ooemachi, Kumamoto City, Japan.

Netherlands

Dr. K. J. van Deen, Laboratorium voor Sociale Geneeskunde, Oostersingel 69 I, Groningen, Netherlands.

WNAR

Mr. Herbert B. Eisenberg, 1329 22nd Street, Santa Monica, California, U. S. A.

Dr. William R. Gaffey, 3119 Eton Avenue, Berkeley 5, California, U. S. A.

Mr. William B. Owen, Statistical Laboratory, Colorado State University, Fort Collins, Colorado, U. S. A.

Mr. Patrick K. Tomlinson, Calif. State Fisheries, 511 Tuna Street, Terminal Island, San Pedro, California, U. S. A.

NEWS AND ANNOUNCEMENTS

Members are invited to transmit to their National or Regional Secretary (if members at large, to the General Secretary) news of appointments, distinctions, or retirements, and announcements of professional interest.

NEWS ABOUT MEMBERS

Victor Chew has accepted a part-time position (starting September 1, 1961) in the Department of Biostatistics of the Johns Hopkins University. He will divide his time between Baltimore, Maryland and Dahlgren, Virginia, where he is a mathematical statistician in the Operations Research Branch of the U. S. Naval Weapons Laboratory.

John J. Gart will spend 1960–61 on a Postdoctoral Research Fellowship at Birkbeck College, University of London, while on leave from his position of Assistant Professor of Biostatistics at The Johns Hopkins University.

David G. Gosslee, formerly with the University of Connecticut, has joined the Statistics Section of the Mathematics Panel at the Oak Ridge National Laboratory where he will consult with biologists.

John Gurland, formerly Professor of the Department of Statistics, Iowa State University, recently accepted a position as Professor of the Mathematics Research Center, U. S. Army, at the University of Wisconsin.

Vincent Hodgson has completed this graduate study at the London School of Economics and Political Science and joined the faculty of The Department of Statistics, The Florida State University, Tallahassee, Florida.

Maurice G. Kendall, professor of statistics at the University of London and president of the Royal Statistical Society, has been appointed to the board of C-E-I-R (U.K.) Ltd. He will assume the new post of director for the London-based company's mathematics, statistics and operations research departments, effective October 1, 1961 and will then vacate his chair at the university.

Eugene Lukacs of the Catholic University of America will take a sabbatical leave during the academic year 1961/62. He will spend the greater part of this time at the Institut Statistique de l'Université de Paris working under an Air Force Grant. From April 1962 to July 1962, he will be Visiting Professor at the Swiss Federal Institute of Technology.

INTERNATIONALES SEMINAR
über
biometrische Methoden in der Medizin und
Genetik

veranstaltet von der Schweizerisch-österreichischen
Gruppe der Internationalen Biometrischen Gesellschaft

vom 18. bis 22. September 1961 in Wien. Österreich.

Der Zweck des Seminars besteht darin, den Teilnehmern eine grundlegende und systematische Ausbildung in der biometrischen Behandlung medizinischtherapeutischer Probleme (Grundprinzipien des Planens und Auswertens einschlägiger Versuche, spezielle Versuchspläne und ihre Begründung, sequentielle Methoden) und Fragen der Genetik (Genfrequenzschätzung, Evolutions- und Mutationsfragen, Vererbung quantitativer Merkmale) zu vermitteln. Beide Themenkreise sind sowohl für die Mediziner als auch die Genetiker aktuell und interessant. Im Sinne einer wirksamen Instruktionsveranstaltung sind täglich höchstens vier Vorlesungen vorgesehen. An mathematischen Grundkenntnissen wird nur der übliche Mittelschullehrstoff vorausgesetzt. Das Lehrprogramm wird von ausgewählten Spezialisten betreut. Da eine begrenzte Teilnehmerzahl vorgesehen ist, ersuchen wir schon jetzt um Vormerkungen beim örtlichen Tagungssekretariat: Institut für Statistik an der Universität Wien, Wien I., Rathausstraße 19 11 3.

Der Versand des detaillierten Programmes und des endgültigen Anmeldungsformulares erfolgt im Sommer 1961. Teilnehmerbeitrag: sFr 30.—(bzw. ö. S 180.—, DM 30.—).

Prof. Dr. A. Linder, Genf
Prof. Dr. H. L. Le Roy, Zürich

Prof. Dr. S. Sagoroff, Wien
Prof. Dr. L. Schmetterer, Wien

BIOMETRIE—PRAXIMETRIE

Tome I, N° 3–4 *Juillet-Octobre, 1960*

Editorial

Utilisation de calculateurs électroniques dans l'analyse d'expériences avec répétitions et de groupes d'expériences de même plan expérimental . F. Yates

Planification des expériences d'alimentation des vaches laitières . . . H. L. Lucas

La sélection artificielle suivant des caractères quantitatifs . . . . . H. L. Le Roy

La répartition des sexes chez des cucurbitacées hybrides et 1 leur relation avec une distribution contagieuse de Neyman . . . . F. Weiling

La puissance du critère F dans l'analyse de la variance de plans en blocs au hasard, nomogrammes pour le choix du nombre de répétitions . . . . M. Keuls

Développement des méthodes biométriques et statistiques dans la recherche agronomique au Congo Belge et au Ruanda-Urundi . . . . J. M. Henry

Développement des méthodes biométriques et statistiques dans la recherche agronomique en Belgique . . . . L. Martin

L'emploi des méthodes biométriques dans l'étude et l'expérimentation des variétés de betterave sucrière en Belgique . . . . M. Simon et N. Roussel

Les méthodes d'échantillonnage en sylviculture . . . . Anne Lenger

Radiosensibilité des graines d'andropogon issues de terrains uranifères et non-uranifères du Katanga . . . . . . Dr. J. Mewissen, J. Damblon, et Z. M. Bacq

Dix années d'activités de la Biometric Society en Belgique, au Congo Belge et au Ruanda-Urundi

Revue bibliographique.

Tome II, N° 1 *Mars, 1961*

Quelques problèmes de numérations bactériennes . . . . H. Grimm

Comparaison de deux méthodes d'analyse de données non orthogonales dans le cas d'expériences à deux facteurs . . . . . . P. Gilbert et B. Grossmann

Analyse de données non orthogonales dans le cas d'une expériences à deux facteurs . . . . A. Rotti

Revue bibliographique: D. J. Finney: Experimental Design and its Statistical Basis . . . . L. Martin

Publishers for the

SOCIETE ADOLPHE QUETELET
7 rue Héger-Bordet, Brussels 1, Belgium

Presses Universitaires de Bruxelles
50, av. Franklin Roosevelt
Brussels

The Biometric Society

BIOMETRICS

FOUNDED BY THE BIOMETRICS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION

TABLE OF CONTENTS

Articles

Queries and Notes

Book Reviews

Number 4 December 1961 Volume 17

BIOMETRICS

Officers of the Biometric Society

General Officers

President: Leopold Martin

Secretary: M. J. R. Healy; *Treasurer*: M. A. Kastenbaum

Council Members

1959–1961	1960–1962	1961–1963
M. S. Bartlett, *BR*	G. S. Watson, *ENAR*	A. W. Kimball, *ENAR*
D. G. Chapman, *WNAR*	C. I. Bliss, *ENAR*	H. N. Turner, *AR*
C. W. Emmens, *AR*	D. J. Finney, *BR*	S. C. Pearce, *BR*
F. G. Fraga, *R. Bras.*	A. Linder, *Switzerland*	J. L. Hodges, *WNAR*
A. Lenger, *R. Belg.*	P. V. Sukhatme, *R. Ital.*	W. U. Behrens, *DR*
C. C. Li, *ENAR*	G. Teissier, *RF*	H. L. LeRoy, *Switzerland*
	F. Yates, *BR*	

Regional Officers

Region	*President*	*Secretary*	*Treasurer*
Australasian	M. Belz	W. B. Hall	G. W. Rogerson
Belgian	M. Welsch	L. Martin	P. Gilbert
Brazilian	A. M. Penha	E. Berquó	A. Groszmann
British	J. A. Fraser Roberts	C. D. Kemp	P. A. Young
E. N. American	Oscar Kempthorne (Henry L. Lucas)	Erwin L. LeClerg	Donald A. Gardiner
French	Ph. L'Heritier	Sully Ledermann	Sully Ledermann
German	O. Heinisch	R. Wette	M. P. Geppert
Italian	G. Montalenti	R. Scossiroli	F. Sella
W. N. American	W. A. Taylor	W. A. Becker	Bernice Brown

National Secretaries

Denmark	N. F. Gjeddebæk	*Netherlands*	H. de Jonge
India	A. R. Roy	*Norway*	L. K. Strand
Japan	M. Hatamura	*Sweden*	H. A. O. Wold
	Switzerland	H. L. LeRoy	

THE POISSON PASCAL DISTRIBUTION[1]

S. K. KATTI[2] AND JOHN GURLAND[3]
Iowa State University of Science and Technology
Ames, Iowa, U. S. A.

1. INTRODUCTION

Elementary distributions such as the Poisson, the Logarithmic and the Binomial which can be formulated on the basis of simple models have been found to be inadequate to describe the situations which occur in a number of phenomena. The Neyman Type A (cf. Evans [5]), the Negative Binomial (cf. Bliss and Fisher [3]), and the Poisson Binomial (cf. McGuire *et al.* [8]), which combine two of the elementary distributions through the processes of compounding and generalizing (cf. Gurland [7]), have been fitted with varying degrees of success to data from a number of biological populations. The aim of this paper is to study what may be called the Poisson Pascal distribution which includes the Neyman Type A and Negative Binomial as particular limiting cases and serves as a natural complement of the Poisson Binomial.

2. FORMULATION OF POISSON PASCAL DISTRIBUTION:

Consider the well-known example of the egg masses and the larvae illustrating the process of generalizing (cf. Neyman [9]). Suppose that the egg masses in a field have a Poisson distribution with probability generating function (p.g.f) exp. $\{\lambda(z - 1)\}$ and that the survivors within an egg mass have a Pascal[4] distribution with p.g.f. $(q - pz)^{-k}$, $p > 0$, $q = 1 + p$, $k > 0$. For the sake of references, we note that, if $g(z)$ is the p.g.f. of a distribution, then the coefficient of z^x in the Taylor's expansion of $g(z)$ yields the probability of count x. Then, by following the arguments in Gurland [7], we obtain for the distribution of the survivors in the field, the Poisson Pascal distribution with p.g.f.

[1] This research was sponsored in part by the Air Force Office of Scientific Research.

[2] Current address: The Florida State University, Tallahassee, Florida.

[3] On leave at Mathematics Research Center, U. S. Army, University of Wisconsin.

[4] Feller [6] refers to the distribution with p.g.f. $(q - pz)^{-k}$, $p > 0$, $q = 1 + p$, $k = 1, 2, 3 \cdots$ etc. However, we will use the work Pascal as synonymous with Negative Binomial and use it in the place of the latter since it is shorter.

$$\exp\{\lambda[(q - pz)^{-k} - 1]\}, \quad \lambda > 0, \quad k > 0, \quad p > 0, \quad q = 1 + p. \tag{1}$$

It is believed that most of the deviation of the distribution of the survivors from the simple distributions mentioned above stems from (i) the complex structure of survivors within an egg mass and (ii) the movement of survivors from place to place. Since the Negative Binomial distribution has been found to be very useful in fitting data involving this type of heterogeneity, it is reasonable to suppose that the Poisson Pascal will give a better description of the population of the survivors in a field.

The Poisson Pascal distribution can also be looked upon as the result of compounding a Pascal distribution with p.g.f. $(q - pz)^{-k_1}$ by taking k_1 to behave as kx where k is a positive constant and x a Poisson random variable with p.g.f. $\exp\{\lambda(z - 1)\}$.

3. SOME PROPERTIES OF THE POISSON PASCAL DISTRIBUTION

The limiting forms that the Poisson Pascal distribution takes as the parameters take on extreme values are given in Table 1. It is to be noted that the Neyman Type A and the Pascal distributions are among the limiting forms.

TABLE 1

SOME LIMITING FORMS OF THE POISSON PASCAL DISTRIBUTION

No.	Limits Taken	Name and p.g.f. of the limit
1	$k \to \infty, p \to 0$ $pk = \lambda_1$	Neyman Type A, $\exp\{\lambda[\exp(\lambda_1(z - 1) - 1]\}$
2	$k \to 0, \lambda \to \infty$ $\lambda k = k_1$	Negative Binomial, $(q - pz)^{-k_1}$
3	$p \to 0, \lambda \to \infty$ $\lambda kp = \lambda_1$	Poisson, $\exp\{\lambda_1(z - 1)\}$

The flexibility of the Poisson Pascal was compared quantitatively with that of the Neyman Types A, B, C, Pascal, and Poisson Binomial by evaluating the relative skewness and kurtosis of each for fixed mean k_1p_1 and variance $k_1p_1(1 + p_1)$ using the indices $\kappa_{[3]}/k_1p_1^3$ and $\kappa_{[4]}/k_1p_1^4$ respectively along the lines of Anscombe [1]. The range of numerical values of these indices for each of the foregoing distributions is shown in Table 2. It is apparent that the Poisson Pascal covers the entire range of distributions from Neyman Type A to Pascal with respect to skewness and kurtosis. Also, the ranges of these ratios for

TABLE 2

Comparison of Skewness and Kurtosis of Certain Distributions

No.	Name	Range of Skewness	Range of Kurtosis
1	Neyman Type A	1	1
2	Neyman Type B	9/8	27/20
3	Neyman Type C	6/5	8/5
4	Pascal	2	6
5	Poisson Pascal	(1,2)	(1,6)
6	Poisson Binomial	(0,1)	(0,1)

the Poisson Pascal and the Poisson Binomial are disjoint. Since their p.g.f.s have the common form

$$\exp \{\lambda[(q - pz)^{-k} - 1]\}, \tag{2}$$

we observe that the distribution with (2) for p.g.f. wherein $\lambda > 0$, $q = 1 + p$, $p > 0$ when $k > 0$ and $-1 < p < 0$ when k is a negative integer, covers a very wide range of distributions. As an aid in computing the values of the ratios for the samples to obtain an idea as to how close the sample distribution is to the various distributions mentioned in Table 2, formulae to compute the factorial cummulants $\kappa_{[i]}$ using sample moments are given in Appendix A.

Since the first two frequencies were large in the sets of data to which this (and similar) distributions were fitted, the ratio of the first two frequencies were compared for some of these distributions. It can be easily shown that the value of this ratio for the Poisson Pascal distribution lies between the ratios for the Neyman Type A and the Pascal distributions. The ratios are given for brevity in Table 3.

TABLE 3

Comparison of the Ratio of the First two Frequencies with Mean and Variance Fixed as k_1P_1 and $k_1P_1(1 + p_1)$

No.	Distribution	Ratio of Frequencies
1	Neyman Type A	$k_1p_1 \exp(-p_1)$
2	Neyman Type B	$2k_1 \{1 - \exp(-3p_1)(1 + 3p_1)\}/(p_1q_1)$
3	Neyman Type C	$3k_1\{2p_1 \exp(-4p_1) + 2[-2 \exp(-4\mathrm{p}_1) + 64p_1^2 + 4p_1]\}/4p_1^3$
4	Pascal	$k_1p_1(1 + p_1)^{-1}$
5	Poisson Pascal	$k_1p_1(1 + p_1)/(k + 1)^{-k-1}$

4. FITTING POISSON PASCAL AND EFFICIENCY OF METHODS OF ESTIMATION

Since obtaining maximum likelihood estimates is very cumbersome, *ad hoc* methods were used to estimate the parameters. When the mean and the variance were large, use was made of the method of the first three moments. When they were moderate and the proportion of the zero frequency large, use was made of the method of the first two moments and the proportion of the zero frequency. When the first two frequencies were large in comparison with the remaining, the method of the first two moments and the ratio of the first two frequencies was used in estimation. The equations for estimation are given in Appencix C. The fit of this distribution to the data of Beall and Rescia [2] are given in Tables 4 and 5. For the sake of reference, the fit of a generalization of the Neyman Type A as given by the authors of these data are also given alongside. The relatively good fit of the Poisson Pascal is apparent.

For obtaining a comparison of the various methods of estimation, it was decided to compute the efficiency function

TABLE 4

Fit of the Observed Frequency of Lespedeza Capitata, Table V of [2]

Plants	Observed Frequency	Expected frequency due to Poisson Pascal (Method of Moments)	Expected frequency as in [2]
0	7178	7185.0	7217.6
1	286	276.0	218.6
2	93	94.5	105.5
3	40	41.5	50.9
4	24	20.2	24.5
5	7	10.4	11.8
6	5	5.6	5.7
7	1	3.1	2.8
8	2	1.7	1.3
9	1	1.0	.6
10	2	.6	.3
11+	1	.3	.4
χ^2	—	9.58	42.97
Degrees of Freedom	—	8	9

E = 1/(Generalized variance × Information determinant)
(cf. Cramer [4], pp. 489–497). (3)

If we denote the parameter vector (λ, p, k) by $(\lambda_1, \lambda_2, \lambda_3)$ and the set of statistics used by (t_1, t_2, t_3), we get the expression for the generalized variance of the estimates as

TABLE 5

Fit of Observed Frequency of Leptinotarsa Decemlineata, Table III of [2]

Insects	Observed	Expected frequency due to Poisson Pascal, (method of two moments and first frequency	Expected frequency as in [2]
0	33	33.0	39.5
1	12	9.8	6.0
2	5	7.4	4.9
3	6	5.5	3.4
4	5	4.0	3.2
5	0	2.9	2.5
6	2	2.1	2.0
7	2	1.5	1.6
8	2	1.1	1.3
9	0	.8	1.0
10	1	.6	.8
11+	2	1.3	3.3
χ^2	—	6.88	13.75
Degrees of Freedom	—	8	9

$$G = \mid V(t_1,t_2,t_3) \mid \Big/ \left| \frac{\partial(\tau_1, \tau_2, \tau_3)}{\partial(\lambda_1, y_2, \lambda_3)} \right|^2, \tag{4}$$

where τ_1, τ_2, τ_3 are the functions of λ_1, λ_2 and λ_3 estimated consistently by t_1, t_2, t_3. A proof of this is given in Appendix B. Since evaluating the covariance matrix $V(t_1, t_2, t_3)$ and the derivatives of t_1, t_2 and t_3 follows from the regular statistical techniques, no elaboration need be made here. A formula for the information determinant (cf. Shenton [10]) is

$$n^3 I = \begin{vmatrix} \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_1} \frac{\partial P_x}{\partial \lambda_1} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_1} \frac{\partial P_x}{\partial \lambda_2} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_1} \frac{\partial P_x}{\partial \lambda_3} \\ \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_2} \frac{\partial P_x}{\partial \lambda_1} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_2} \frac{\partial P_x}{\partial \lambda_2} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_2} \frac{\partial P_x}{\partial \lambda_3} \\ \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_3} \frac{\partial P_x}{\partial \lambda_1} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_3} \frac{\partial P_x}{\partial \lambda_2} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_3} \frac{\partial P_x}{\partial \lambda_3} \end{vmatrix}. \tag{5}$$

The principal problems in evaluating I for a value of (λ, p, k) therefore are *(i)* to evaluate the various P_x and the derivatives of P_x and *(ii)* to determine the number of terms to be used in summing the infinite series. To obtain P_x, let

$$g(z) = \exp \{\lambda[(q - pz)^{-k} - 1]\} \tag{6}$$

and

$$h(z) = (q - pz)^{-k}.$$

By differentiating (6) successively we get

$$g'(z) = \lambda g(z)h'(z), \tag{7}$$

and

$$g^{(x+1)}(z) = \lambda \sum_{i=0}^{x} \binom{x+1}{i} g^{(i)}(z)h^{(x+1-i)}(z). \tag{8}$$

Set $z = 0$ in equations (6), (7) and (8) and observe that $g^{(x)}(0) = x!\, P_x$ and $h^{(x)}(0) = x!\, \pi_x$ where

$$\pi_x = \frac{(k + x - 1)!}{(k - 1)!\, x!} p^x q^{-k-x}, \tag{9}$$

is the probability of x in the Pascal distribution with $h(z)$ as p.g.f. Then we have the recurrence formulae

$$P_0 = \exp. \{\lambda[(q)^{-k} - 1]\} \tag{10}$$

and

$$P_{x+1} = \frac{\lambda}{x + 1} \left\{ \sum_{i=0}^{x} (x + 1 - i)\pi_{x+1} - i\, P_i \right\}, \tag{11}$$

which can be repeatedly used to evaluate any P_x.

The various derivative are given by

$$\frac{\partial P_x}{\partial \lambda_1} = \frac{\partial P_x}{\partial \lambda} = \left\{ \frac{\partial^x}{\partial z^x} \left[\frac{\partial g(z)}{\partial \lambda} \right] \right\}_{z=0} = \frac{1}{x!} \left\{ \frac{\partial^x}{\partial z^x} [g(z)(h(z) - 1)] \right\}_{z=0}$$

$$= \frac{q}{\lambda k p} (x + 1)P_{x+1} - P_r \left(1 + \frac{r}{\lambda k}\right), \tag{12}$$

$$\frac{\partial P_x}{\partial \lambda_2} = \frac{\partial P_x}{\partial p} = \frac{1}{x!}\left\{\frac{\partial^x}{\partial z^x}\left[\frac{\partial}{\partial P}\, g(z)\right]\right\}_{z=0} = \frac{1}{x!}\left\{\frac{\partial^x}{\partial z^x}\left[\frac{z}{p}\, g'(z)\right]\right\}_{z=0}$$

$$= \frac{1}{p}\,\{xP_x - (x+1)P_{x+1}\} \qquad (13)$$

and

$$\frac{\partial P_x}{\partial \lambda_3} = \frac{\partial P_x}{\partial k} = \frac{1}{x!}\left\{\frac{\partial^x}{\partial z^x}\left[\frac{\partial}{\partial r}\, g(z)\right]\right\}_{z=0}$$

$$= \frac{1}{x!}\left\{\frac{\partial^x}{\partial z^x}\,[-\lambda g(z) h(z) \log (q - pz)]\right\}_{z=0}$$

$$= \frac{1}{kp}\sum_{i=1}^{x} (p/q)^i\, \frac{1}{i}\,\{q(x - i + 1)P_{x-i+1} - p(x - i)P_{x-i}\}$$

$$- \frac{1}{kp}\,(\log q)\{q(x+1)P_{x+1} - pxP_x\} \qquad (14)$$

for all x.

TABLE 6

EFFICIENCY OF THE METHOD OF THE FIRST THREE MOMENTS FOR THE POISSON PASCAL

				k		
λ	p	.1	.3	.5	1.0	2.0
.1	.1	.84	.82	.82	.81	.76
.1	.3	.59	.58	.58	.54	.47
.1	.5	.45	.44	.43	.40	.33
.1	1.0	.26	.25	.24	.22	.18
.1	2.0	.13	.12	.12	.12	.13
.5	.1	.90	.81	.82	.77	.67
.5	.3	.59	.58	.56	.49	.35
.5	.5	.46	.44	.41	.34	.22
.5	1.0	.26	.25	.23	.18	.10
.5	2.0	.13	.13	.12	—	—
1.0	.1	.81	.83	.82	.75	.63
1.0	.3	.59	.59	.55	.46	.28
1.0	.5	.46	.44	.40	.31	.15
1.0	1.0	.27	.25	.23	.15	.05
1.0	—	—	—	—	—	—
5.0	.1	.94*	.99*	.78*	.58*	.21*
5.0	.3	.62*	.56*	.41*	.11*	.01*
5.0	.5	.48*	.38*	.20*	.04*	.00*
5.0	1.0	.29*	.17*	.06*	.02*	.00*
5.0	—	—	—	—	—	—

To determine the number of terms we use the following rule:

Let $T_{ij}(n)$ denote the $(n + 1)$th term in the series involved in the (i, j)th term of matrix (5). Let $S_{ij}(n) = \sum_{r=0}^{n} T_{ij}(r)$. Compute $S_{ij}(n)$ and $\sum_{ij} T^2_{ij}(n)/S^2_{ij}(n)$ for $n = 1, 2, \cdots$ *et cetera* successively till a value of n is reached for which

$$\sum_{ij} T^2_{ij}(n)/S^2_{ij}(n) < 10^{-8}. \qquad (15)$$

It is clear from (15) that $| T_{ij}(n) |/| S_{ij}(n) | < 10^{-4}$ for each i and j. If the series converge faster than a geometric series with common ratio less than 0.9 and this convergence starts before the value of n is reached for which (15) is satisfied, the calculated efficiency will be correct to three significant figures. If the significant figures do not cancel out, this should yield the efficiencies computed therefrom, correct to three decimal places. When the inequality (15) was not satisfied for values of $n \leq 20$, the partial sum of the first twenty terms was taken as the value of the series since evaluating the terms of the series

TABLE 7

EFFICIENCY OF THE METHOD OF THE FIRST TWO MOMENTS AND THE FIRST FREQUENCY FOR THE POISSON PASCAL DISTRIBUTION AT λ = 0.1

				k		
λ	p	.1	.3	.5	1.0	2.0
.1	.1	.99	.98	.98	.98	.98
.1	.3	.93	.94	.95	.94	.93
.1	.5	.90	.90	.91	.91	.89
.1	1.0	.82	.83	.84	.84	.83
.1	2.0	.74	.76	.78	.84	.78
.5	.1	1.00	1.00	.98	.96	.95
.5	.3	.93	.94	.93	.92	.87
.5	.5	.91	.91	.90	.89	.82
.5	1.0	.83	.85	.85	.82	.73
.5	2.0	.75	.78	.81	—	—
1.0	.1	1.00	.99	.97	.96	.94
1.0	.3	.94	.96	.94	.93	.87
1.0	.5	.92	.92	.91	.98	.81
1.0	1.0	.84	.85	.86	.82	.72
1.0	2.0	.88*	.87*	.87*	.87*	.83*
5.0	.1	—	—	.99*	.89*	.98*
5.0	.3	.97*	.99*	1.00*	1.00*	.98*
5.0	.5	.95*	.97*	.98*	.99*	.96*
5.0	1.0	.90*	.96*	.97*	—	—
5.0	2.0	.89*	—	—	—	—

for n larger than 20 is very time consuming. The efficiency when n was restricted to 20 is marked with an asterisk to indicate that they are less likely to be correct to three decimal places. When the efficiency so computed was larger than one (due to the inaccuracy in computing the information determinant), the corresponding cell in the efficiency table is left blank.

The efficiency of the method of moments is given for certain values of (λ, p, k) in Table 6. The efficiency of the method of the first two moments and the first frequency is given in Table 7 and that of the method of the first two moments and the ratio of the first two frequencies in Table 8 for the same values of (λ, p, k).

It is apparent that the method of the first two moments and the ratio of the first two frequencies has high efficiency and is superior to the other two methods when λ is small (and consequently the first two counts account for a large proportion of the observed frequencies). Also the method of the first two moments and the first frequency is highly efficient when λ is moderately large. When λ, p or k approaches

TABLE 8

EFFICIENCY OF THE METHOD OF THE FIRST TWO MOMENTS AND THE RATIO OF THE FIRST TWO FREQUENCIES FOR THE POISSON PASCAL AT $\lambda = 0.1$

				k		
λ	p	.1	.3	.5	1.0	2.0
.1	.1	1.00	1.00	1.00	1.00	1.00
.1	.3	.98	.99	.99	.99	.98
.1	.5	.99	.98	.98	.97	.94
.1	1.0	.94	.94	.93	.91	.83
.1	2.0	.87	.88	.89	.91	.99
.5	.1	1.00	.99	1.00*	1.00	.99
.5	.3	.98	1.00	.99	.98	.96
.5	1.0	.99	.98	.97	.96	.92
.5	2.0	.94	.94	.93	.89*	.80*
1.0	.1	.98	1.00	1.00	1.00	.99
1.0	.3	.99	1.00	.99	.99	.95
1.0	.5	1.00	.99	.98	.96	.91
1.0	1.0	.95	.94	.93	.88	.79
1.0	2.0	.88*	.89*	.91*	—	—
5.0	.1	—	—	.99*	.95*	.91*
5.0	.3	1.00	.99*	.98*	.89*	.76*
5.0	.5	1.00	.96*	.91*	.80*	.65*
5.0	1.0	.95*	.88*	.79*	.84*	1.00*
5.0	2.0	.90*	.98*	—	—	—

infinity, it can be shown by calculus that the method of moments is much superior to the other two but since the efficiency of each of these methods tends to zero for such values, this has little significance.

5. CONCLUSIONS

On the basis of the properties discussed in Section 3 and the fitting in Section 4, we observe that the Poisson Pascal distribution acts as a bridge between the Neyman Type A and the Negative Binomial distributions and may be used with advantage when the latter distributions are inadequate to represent the population accurately.

From the tables of efficiency, it is clear that in the region of tabulations, at least one of the *ad hoc* methods of estimation suggested above has high efficiency. It is believed that in practice, (λ, p, k) will not be far beyond the region of tabulation and that one of these methods can be used without too much loss of information. Techniques for choosing one of the many *ad hoc* methods on the basis of the sample will be discussed in a future paper.

REFERENCES

1. Anscombe, F. J. [1950]. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika 37*, 358–82.
2. Beal, G. and Rescia, R. [1953]. A generalization of Neyman's contagious distribution. *Biometrics 9*, 354–86.
3. Bliss, C. I. and Fisher, R. A. [1958]. Fitting of negative binomial distribution to biological data. *Biometrics 9*, 176–200.
4. Cramer, H. [1946]. *Mathematical Methods of Statistics*, Princeton Univ. Press, Princeton, N. J.
5. Evans, D. A. [1953]. Experimental evidence concerning contagious distributions in ecology. *Biometrika 40*, 186–210.
6. Feller, William [1957]. *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, New York, N. Y.
7. Gurland, J. [1957]. Some interrelations among compound and generalized distributions. *Biometrika 44*, 265–68.
8. McGuire, J. U., Brindley, T. A. and Bancroft, T. A. [1957]. The distribution of European corn-borer larvae *Pyrausta Nubilalis* (HBN), in field corn. *Biometrics 13*, 65–78.
9. Neyman, J. [1939]. On a new class of "contagious" distributions applicable in entomology and bacteriology. *Ann. Math. Stat. 10*, 35–57.
10. Shenton, L. R. [1949]. On the efficiency of the method of moments and Neyman's Type A distribution. *Biometrika 36*, 450–54.

APPENDIX

A. Formulae to compute factorial cumulants using moments about the origin:

We first obtain formulae to compute the first four factorial moments $\mu_{[i]}$, $i = 1, \cdots, 4$ using moments about the origin μ_i and then obtain formulae to evaluate the first four factorial cumulants $\kappa_{[i]}$ $i = 1, \cdots, 4$ using these factorial moments.

As for the first objective we note that $\mu_{[i]} = E\{x(x-1)\cdots(x-i+1)\}$, $i = 1, 2, \cdots$. By expanding the product within the expectation sign and using the elementary properties of the expectation operator, we get

$$\mu_{[1]} = \mu_1', \qquad \mu_{[2]} = \mu_2' - \mu_1', \qquad \mu_{[3]} = \mu_3' - 3\mu_2' + 2\mu_1',$$

and

$$\mu_{[4]} = \mu_4' - 6\mu_3' + 11\mu_2' - 6\mu_1'. \tag{16}$$

As for the latter, we observe that if $u(t)$ and $\psi(t)$ denote the factorial moment generating function and the factorial cumulant generating function, then $\psi(t) = \log u(t)$. On differentiating the equation successively with respect to t at $t = 0$ and noting that $\kappa_{[i]} = \psi^{(i)}(0)$, we have

$$\kappa_{[1]} = \mu_{[1]}, \qquad \kappa_{[2]} = \mu_{[2]} - \mu_{[1]}^2,$$
$$\kappa_{[3]} = \mu_{[3]} - 3\mu_{[1]}\mu_{[2]} + 2\mu_{[1]}^3,$$

and

$$\kappa_{[4]} = \mu_{[4]} - 3\mu_{[2]}^2 - 4\mu_{[1]}\mu_{[3]} + 12\mu_{[1]}^2\mu_{[2]} - 6\mu_{[1]}^4. \tag{17}$$

B. To prove formula (4):

Since τ_1, τ_2, τ_3 are functions of λ_1, λ_2 and λ_3, let us write them more explicitly as $\tau_1(\lambda_1, \lambda_2, \lambda_3)$, $\tau_2(\lambda_1, \lambda_2, \lambda_3)$ and $\tau_3(\lambda_1, \lambda_2, \lambda_3)$ respectively. If $\hat\lambda_1, \hat\lambda_2, \hat\lambda_3$ are the estimates of $\lambda_1, \lambda_2, \lambda_3$, then using the statistics t_1, t_2, t_3 which estimate the τ's consistently, we have

$$t_i = \tau_i(\hat\lambda_1, \hat\lambda_2, \hat\lambda_3), \qquad i = 1, 2, 3.$$

Hence, we have the asymptotic relations

$$t_i - \tau_i = \left(\frac{\partial\tau_i}{\partial\lambda_1}, \frac{\partial\tau_i}{\partial\lambda_2}, \frac{\partial\tau_i}{\partial\lambda_3}\right)(\hat\lambda_1 - \lambda_1, \hat\lambda_2 - \lambda_2, \hat\lambda_3 - \lambda_3)', \quad i = 1, 2, 3.$$

which can be rewritten as

$$\begin{bmatrix} \hat\lambda_1 - \lambda_1 \\ \hat\lambda_2 - \lambda_2 \\ \hat\lambda_3 - \lambda_3 \end{bmatrix} = \left[\frac{\partial(\tau_1, \tau_2, \tau_3)}{\partial(\lambda_1, \lambda_2, \lambda_3)}\right]^{-1} \begin{bmatrix} t_1 - \tau_1 \\ t_2 - \tau_2 \\ t_3 - \tau_3 \end{bmatrix}.$$

The generalized variance G of $\hat\lambda_1, \hat\lambda_2, \hat\lambda_3$ is then given by

$$G = | E(\lambda_1 - \lambda_1, \lambda_2 - \lambda_2, \lambda_3 - \lambda_3)'(\hat\lambda_1 - \lambda_1, \hat\lambda_2 - \lambda_2, \hat\lambda_3 - \lambda_3) |$$
$$= | V(t_1, t_2, t_3) | \Big/ \left| \frac{\partial(\tau_1, \tau_2, \tau_3)}{\partial(\lambda_1, \lambda_2, \lambda_3)} \right|^2$$

which is formula (4).

C. Equations for estimation:

We give below the equations for estimating the parameters for the various methods mentioned in Section 4. Their derivations are omitted for brevity.

i. Equations for the estimation of parameters using the first three moments are:

$$k = \frac{\hat{\kappa}_{[3]}\hat{\kappa}_{[1]}}{\hat{\kappa}_{[3]}\hat{\kappa}_{[1]} - (\hat{\kappa}_{[2]})^2} - 2, \tag{18}$$

$$p = \hat{\kappa}_{[2]}/(\hat{\kappa}_{[1]}(k + 1)), \tag{19}$$

$$\lambda = \hat{\kappa}_{[1]}/(kp). \tag{20}$$

ii. Equations for the estimation of parameters using the first two moments and the proportion $\hat{P}_0$ of zeros are:

$$p \log \left\{1 + \left(\frac{\hat{\kappa}_{[2]}}{\hat{\kappa}_{[1]}} - p\right) \frac{\log \hat{P}_0}{\hat{\kappa}_{[1]}}\right\} + \left(\frac{\hat{\kappa}_{[2]}}{\hat{\kappa}_{[1]}} - p\right) \log (1 + p), \tag{21}$$

$$k = \frac{\hat{\kappa}_{[2]}}{p\hat{\kappa}_{[1]}} - 1, \qquad \lambda = \hat{\kappa}_{[1]}/(kp). \tag{22}$$

iii. Equation for the estimation of the parameter p using the first two moments and the ratio $\hat{P}_1/\hat{P}_0$ of the first two frequencies is

$$\frac{\log (1 + p)}{p} = \frac{\hat{\kappa}_{[1]}}{\hat{\kappa}_{[2]}} \log \{\hat{\kappa}_{[1]}\hat{P}_0/\hat{P}_1\}. \tag{24}$$

k and λ are then estimated by using equations (22) and (23).

SOME RANK SUM MULTIPLE COMPARISONS TESTS

ROBERT G. D. STEEL
Institute of Statistics, North Carolina State College
Raleigh, North Carolina, U. S. A.

1. SUMMARY

Rank sum tests for multiple comparisons and suitable to the completely random design with equal sample sizes are discussed. Significance tables and some facts about the joint distribution of rank sums are given. An example illustrating test procedures and making use of the tables is presented.

2. INTRODUCTION

A number of nonparametric tests based on ranks have been proposed for the comparison of treatments in a completely random design. For example, we have the Wilcoxon-Mann-Whitney test [21, 10], basically a two-sample test with a per-comparison error rate.

Also, Kruskal and Wallis [9] have proposed a rank test which is an analogue of Snedecor's F-test. This test provides evidence concerning the presence of real differences but is of limited use in locating them.

Steel [16, 17] has presented rank tests for comparing treatments against control and for all pairwise comparisons. Both of these tests use experiment-wise error rates.

Pfanzagl [13], as part of a more general theory, has discussed a two-step nonparametric decision process based on ranks, for testing the null hypothesis that k samples come from the same population and, if this is rejected, for deciding which one of the samples comes from a different population. No tables are given but it is suggested that they might be obtained by random sampling. It is also shown that the limiting distribution of the multivariate criterion is multinormal.

The per-comparison error rate test is sometimes criticized, particularly when all possible paired comparisons are made, because it will almost certainly lead to false declarations of significance when the experiment includes many treatments and if customary significance levels are used. It is also deemed inappropriate when the experiment is considered to be the conceptual unit.

The experimentwise error rate test is sometimes criticized because it requires such a large difference to be declared significant that it becomes difficult to detect any but the largest real differences when customary significance levels are used. Also, it may be that the individual comparison is considered to be the conceptual unit.

A brief discussion of these error rates is given by Steel [18] in response to *Biometrics* Query 163.

Choice of a definition of error rate in the conduct of a particular experiment seems somewhat less crucial when it is realized that we can compute the significance level for a particular definition of error rate from knowledge of the chosen significance level for any other definition of error rate. This is not generally a simple computation unless the comparisons are independent. In the case of p independent comparisons, if α' and α are the experimentwise and per-comparison error rates respectively, we have the relation: $1 - \alpha' = (1 - \alpha)^p$.

When comparisons are not independent, computation of comparable significance levels for different definitions of error rate depends upon the extent of the dependence and the nature of the multiple comparisons test. No individual is likely to perform such a computation for a single experiment. Thus a table needs to be prepared for comparing the customary significance levels for differently defined error rates. For tests based on an underlying normal distribution, this has been done fairly extensively by Harter [4, 6].

The experimenter may try to meet the usual criticisms of per-comparison and experimentwise error rates by choice of a non-standard significance level or an alternative test procedure. Presently, tables of significant values for such levels do not appear to be available for experimentwise error rates; in the case of alternative tests, several are available, including the Newman-Keuls [11, 7] procedure and Duncan's [1,5] new multiple range test, which are sequential in nature.

This paper is concerned with rank tests, in particular with tables for an all-pairs-of-treatments test with an experimentwise error rate, and the use of these tables for a fixed rank sum test, an analogue of Tukey's w-procedure (for example, see Steel and Torrie [19]), and for two multiple rank sum tests, analogues of the Newman-Keuls procedure and of Duncan's test. Table 2 is used in the first two cases, Table 3 in the last case.

3. CONSTRUCTION OF TABLES

The proposed tests call for rank sums and their conjugates computed for all pairwise comparisons of treatments. The minimum of each rank sum and conjugate is used, the set of minima providing a multivariate rank sum test criterion. These sums are compared with a single tabulated value for the fixed rank sum test and with several values for the multiple rank sum tests. Table 2 provides critical values for the analogues of Tukey's and the Newman-Keuls tests; Table 3 provides for the analogue of Duncan's test.

Methods for constructing probability tables and limited tables have been presented earlier [16, 17]. Construction of exact tables of any extent was beyond the computing facilities available. However, some machine time was available and this was used for some sampling experiments.

It was originally intended to ignore the discrete nature of the data and to use the Kolmogorov-Smirnov [8, 14] one-sample test to determine the sample size necessary to attain a certain precision in the constructed tables. However, available computing facilities limited sampling to values of $k = 3$ and 4 (2 and 3 treatments when one was control) and $n = 4$ (1) 10. In addition, samples were obtained for $k = 5$, $n = 4$, 5 and $k = 6$, $n = 5$. The number of permutations obtained for each case was either 5000 or 6000. These tables were used only for checking purposes against the few exact distributions available, $k = 3$ and $n = 3$ (1) 6, and against approximations used in constructing these and earlier tables.

It was assumed that the various multivariate rank sum criteria are distributed approximately as multinormal distributions having mean vectors and variance-covariance matrices as given in the appendix. (Fraser's [3] vector form of the Wald-Wolfowitz-Hoeffding theorem does not apply since the $\| C_{n\alpha}(i, j) \|$'s of Fraser do not exist for the test criteria used here.)

On this assumption, one naturally proceeds to base computation on presently available tables. Tables for Tukey's and Duncan's tests are the obvious choice for all-pairs tests. These tests are based on a multinormal distribution with $\rho^2 = n_i n_j/(n_h + n_i)(n_h + n_j)$ where the present distribution calls for $\rho^2 = n_i n_j/(n_h + n_i + 1)(n_h + n_j + 1)$, a small difference. The appropriate tables are, then, tables of the Studentized range with known variance, that is, infinite degrees of freedom. Table 22 of Pearson and Hartley [12] is such a table, gives percentage points of .10, .05 and .01, is appropriate for the first two tests, and was used in computing Table 2. Corrected tables for Duncan's test have been computed by Harter [5] and this table was used in computing Table 3, also for percentage points of .10, .05 and .01.

Table 2 was constructed by taking the integral part of $\mu - t\sigma/\sqrt{2}$, unless the decimal fraction was $> .9$, in which case the next higher integer was tabulated, where t was obtained from the distribution of w/σ, w = range, Table 22 of Pearson and Hartley [12]. Since rank sums are essentially differences, it is necessary to introduce $\sqrt{2}$ into the denominator as shown. Tabulated rank sum values for $\alpha = .10$ differed in only two cases from values obtained by sampling. In particular, for $k = 4$, $n = 4$, no value is significant by sampling; for $k = 3$, $n = 6$, a rank sum of 25 is significant by sampling whereas 26 is not. Values for $\alpha = .05$ differed in no case. Values for $\alpha = .01$ ran lower than those obtained by sampling, the difference increasing with n to a value of two in three cases. Hence, it is reasonable to conclude that tabulated values of the rank sum are conservative (low) for $\alpha = .01$.

The first attempt to construct Table 3 led to values which tended to run high for $\alpha = .10$, correct for $\alpha = .05$, and low for $\alpha = .01$, relative to values found by sampling. For this reason, tabulated values are of $\mu - t\sigma/\sqrt{2}$ decreased by unity for $\alpha = .10$, as computed for $\alpha = .05$, and increased by unity for $\alpha = .01$. On this basis, tabulated values for $\alpha = .10$ appear to be low, hence conservative, when not in agreement with sampling results; in particular, 7 out of 19 tabulated values are one unit low. For $\alpha = .05$, 3 out of 19 values are one unit high. For $\alpha = .01$, 5 out of 19 values are one unit low, with three of these being for $n = 5$.

Tables have already been constructed [16], using Dunnett's [2] tables, for the treatments against control test. These tables agree well with the sampling results. In no case is there a difference of more than one in the value of the test criterion, with the tables most often giving the conservative (lower) value.

TABLE 1

FINAL WEIGHTS OF CHICKS AT SIX WEEKS (GRAMS) FOR VARIOUS SOURCES OF PROTEIN SUPPLEMENT

H Horse-bean	*L* Linseed Oil Meal	*Sb* Soybean Oil Meal	*Sf* Sunflower Seed Oil Meal	*M* Meat Meal	*C* Casein
179	309	243	423	325	368
160	229	230	340	257	390
136	181	248	392	303	379
227	141	327	339	315	260
217	260	329	341	380	404
168	203	250	226	153	318
108	148	193	320	263	352
124	169	271	295	242	359
143	213	316	334	206	216
140	257	267	322	344	222

4. USE OF TABLES

To illustrate the use of the tables, the data in Query 60 of *Biometrics* (15) are used. These are presented in Table 1. Since the test is presently unavailable for unequal sample sizes, only the first ten items in each treatment are used.

The following set of minimum rank sums is obtained by pairwise rankings, a minimum being the lesser of the rank sum T, and its conjugate, $T' = (2n + 1)n - T$; minimum treatment is the treatment for which the rank sum is minimum. Ties were assigned their average rank. This gives a multivariate criterion with 15 entries.

Comparison	*H, Sf*	*H, Sb*	*H, C*	*L, Sf*	*H, M*	*L, C*
Minimum Rank Sum	56	57	58	60	62	$64\frac{1}{2}$
Minimum Treatment	H	H	H	L	H	L

Comparison	*Sb, Sf*	*H, L*	*L, Sb*	*L, M*	*Sb, C*	*Sf, M*	*M, C*	*Sb, M*	*Sf, C*
Minimum Rank Sum	71	75	75	$75\frac{1}{2}$	80	82	84	100	103
Minimum Treatment	*Sb*	*H*	*L*	*L*	*Sb*	*Sf*	*M*	*Sb*	*Sf*

From Table 2, a rank sum of 67 is significant at the 5 percent level, $k = 6$, $n = 10$; hence six comparisons are declared significant.

The device, used with multiple comparisons procedures, of underlining treatments which cannot be distinguished by their means may be adapted to apply to rank sum procedures. Thus, from the test, it appears that H and L, and L, Sb, M, Sf and C form two groups as a first step; since L can be distinguished from Sf and C, the latter group becomes two, namely L, Sb and M, and Sb, M, Sf and C. We have:

$$\begin{array}{cccccc} H & L & Sb & M & Sf & C \\ \cline{1-2} \\[-2ex] \cline{2-4} \\[-2ex] & & \cline{3-6} \end{array}$$

Ordering of Sb and M, and of Sf and C was done on the basis of rank sums for these paired comparisons though this does not imply that this is the only, or even the best, method.

This procedure is an analogue of Tukey's w-procedure [19]. The significance level is for an experimentwise error rate. Use of Wilcoxon's [21] two-sample test, with its per-comparison error rate, calls for a significant rank sum of 78; this will result in four more comparisons being declared significant.

TABLE 2
Percentage Points of the Minimum Rank Sum
(an approximation)

Number in treatment	α	k = number of treatments being tested							
		3	4	5	6	7	8	9	10
4	.10	10	10	—	—	—	—	—	—
	.05	—	—	—	—	—	—	—	—
	.01	—	—	—	—	—	—	—	—
5	.10	17	16	15	15	—	—	—	—
	.05	16	15	—	—	—	—	—	—
	.01	—	—	—	—	—	—	—	—
6	.10	26	24	23	22	22	21	21	—
	.05	24	23	22	21	—	—	—	—
	.01	—	—	—	—	—	—	—	—
7	.10	36	34	33	32	31	30	30	29
	.05	34	32	31	30	29	28	28	—
	.01	29	28	—	—	—	—	—	—
8	.10	48	46	44	43	42	41	40	40
	.05	45	43	42	40	40	39	38	38
	.01	40	38	37	36	—	—	—	—
9	.10	62	59	57	56	55	54	53	52
	.05	59	56	54	53	52	51	50	49
	.01	52	50	48	47	46	45	—	—
10	.10	77	74	72	70	69	68	67	66
	.05	74	71	68	67	66	64	63	63
	.01	66	63	62	60	59	58	57	56

It is also possible to propose and carry out a sequential procedure, an analogue of the Newman-Keuls procedure, which uses several rank sums for testing. For this procedure, the above analysis is the first step and has separated the treatments into three groups. To proceed, we assume that declared differences are indeed real. Hence to test H versus L, the first group, we may use Wilcoxon's [21] two-sample test, H and L are declared significantly different and the line beneath them may be removed.

Further, compare L, Sb and M using the critical value for $k = 3$, $n = 10$, namely 74. L versus Sb and L versus M at 75 and $75\frac{1}{2}$ are beyond the 10 percent point but are not quite significant. We cannot distinguish among these three treatments.

Finally, consider the group composed of treatments *Sb*, *M*, *Sf* and *C*. The critical value is 71, $k = 4$, $n = 10$. The treatments *Sb* and *Sf* can be distinguished and we must, then, change the order of *Sf* and *C* from that proposed when the fixed rank sum test was used. No further differences will be declared significant by this procedure. We have:

H *L* *Sb* *M* *C* *Sf*

For the Tukey and Newman-Keuls parametric procedures, we find means of 160.2(*H*), 211.0(*L*), 267.4(*Sb*), 278.8(*M*), 323.2(*Sf*) and 326.8(*C*). Also $s_{\bar{x}} = 18.03$ and significant ranges are 51.2, 61.5, 67.6, 71.9 and 75.2 for $k = 2, \cdots, 6$ respectively.

For Tukey's test, the fixed rank sum is 75.2. We find:

H *L* *Sb* *M* *Sf* *C*

For the Newman-Keuls procedure, we find:

H *L* *Sb* *M* *Sf* *C*

We now compare the results obtained from applying the parametric and non-parametric procedures.

The fixed rank sum test and Tukey's test lead to the same conclusions. Both are based on experimentwise error rates.

Conclusions drawn from the multiple rank sum test and the Newman-Keuls test differ as follows. *L* versus *M* is significant by the Newman-Keuls procedure only; *Sb* versus *Sf* is significant by the rank sum test only. Otherwise, the procedures lead to the same conclusions. Fifteen paired tests have been made. Since *L* versus *M* is nearly significant and *Sb* versus *Sf* is just significant by the multiple rank sum test, it would appear that the two methods lead to conlcusions, for this example, that differ only slightly.

The other multiple rank sum test to be considered is an analogue of Duncan's new multiple range test. Table 3 provides critical values. A rank sum of 75 is significant at the 5 percent level, $k = 6$, $n = 10$; hence nine comparisons are significant. Tentatively, we have:

H *L* *M* *Sb* *C* *Sf*

Unfortunately, this includes an anomaly since *Sb* versus *Sf* is also declared significant. The same would be true if the Wilcoxon-Mann-Whitney test were being used at a significance level (between 5 percent and 1 percent) calling for a critical value of 75.

TABLE 3

Percentage Points of the Minimum Rank Sum for Duncan Analogue (an approximation)

Number in treatment	α	k = number of treatments being tested							
		3	4	5	6	7	8	9	10
4	.10	11	10	10	10	10	10	10	10
	.05	10	10	10	10	10	10	10	10
	.01	—	—	—	—	—	—	—	—
5	.10	18	18	17	17	17	17	17	17
	.05	17	17	16	16	16	16	16	16
	.01	15	15	15	—	—	—	—	—
6	.10	27	26	26	26	26	25	25	25
	.05	26	25	25	25	25	24	24	24
	.01	23	22	22	22	22	21	21	21
7	.10	37	37	37	36	36	36	36	36
	.05	36	35	35	35	34	34	34	34
	.01	32	32	31	31	30	30	30	30
8	.10	50	49	49	49	48	48	48	48
	.05	48	47	47	46	46	46	46	45
	.01	43	42	42	41	41	41	41	40
9	.10	65	64	63	63	62	62	62	62
	.05	62	61	60	60	60	59	59	59
	.01	56	55	54	54	53	53	53	52
10	.10	81	80	79	79	78	78	78	77
	.05	77	76	76	75	75	74	74	74
	.01	70	69	68	68	67	67	67	66

On the basis that treatments declared significantly different are indeed so, we proceed to test treatments M, Sb, C and Sf using $k = 4$ and L and M using $k = 2$. The final result is:

$$H \qquad L \qquad Sb \qquad M \qquad C \qquad Sf$$

This result is the same as that obtained using the Wilcoxon-Mann-Whitney test.

Using Duncan's (parametric) new multiple range test, we find:

$$H \qquad L \qquad Sb \qquad M \qquad Sf \qquad C$$

Conclusions from Duncan's test and its rank sum analogue differ only in that the parametric test finds *Sb* versus *C* to be significant.

The multiple rank sum test can be adapted to apply to testing treatments versus control as well.

5. APPENDIX—THEORY

The problem is concerned with pairwise testing of treatments in the one-way classification or completely random design.

The test criteria are rank sums, computed as for Wilcoxon's [21] two sample test, for appropriate pairs of treatments. Rank sums will be referred to Tables 2 or 3 for testing rather than to White's [20] table for the Wilcoxon-Mann-Whitney test.

Two tests will be considered:

1. The all pairs test, in detail.
2. Treatments against control, rather briefly.

For the all pairs test, let X_i, $i = 1, \cdots, k$ be random variables measuring some characteristic for each of k samples or treatments. Let there be n_i observations on the i-th treatment. Computation of the test criterion requires us to:

1. Rank the X_i's and the X_j's, all $i < j$, assigning rank 1 to the least observation.
2. Add ranks for the variable with fewer observations to give T_{ij}. (There is no loss of generality if we assume $n_1 \leq n_2 \leq \cdots \leq n_k$).
3. Compute the conjugate of T_{ij}, namely $T'_{ij} = (n_i + n_j + 1)n_i - T_{ij}$.

The conjugate is the rank total that would be obtained if rank 1 were assigned to the highest observation. Conjugates are required for two-tailed tests.

Consider $(T_{12}, \cdots, T_{1k}, T_{23}, \cdots, T_{k-1,k})$, a criterion with $\binom{k}{2}$ components. Rank tests are based on the assumption that, under H_0, all permutations of the $\sum n_i$ observations are equally likely. Hence, we must know the number of ways in which $(T_{12}, \cdots, T_{k-1,k})$ can be obtained. For this, a recursion formula is given in [14]. This provides a method, though tedious, of deriving the distribution of $(T_{12}, \cdots, T_{k-1,k})$.

The distribution of $(T_{12}, \cdots, T_{k-1,k})$ has been shown to have the following parameters [14]:

$$E(T_{ij}) = \mu_{ij} = n_i(n_i + n_j + 1)/2,$$

$$E(T_{ij} - \mu_{ij})^2 = \sigma^2_{ij} = n_i n_j (n_i + n_j + 1)/12,$$

$$E(T_{hi} - \mu_{hi})(T_{hj} - \mu_{hj}) = \sigma_{hi,hj} = n_h n_i n_j/12 = \sigma_{hj,ij},$$

$$E(T_{hi} - \mu_{hi})(T_{ij} - \mu_{ij}) = \sigma_{hi,ij} = -n_h n_i n_j/12,$$

$$E(T_{gh} - \mu_{gh})(T_{ij} - \mu_{ij}) = \sigma_{gh,ij} = 0,$$
$$\rho^2_{ih,hj} = \rho^2_{ih,jh} = \rho^2_{hi,hj} = n_i n_j/(n_h + n_i + 1)(n_h + n_j + 1),$$
$$\rho^2_{gh,ij} = 0.$$

The determinant of the variance-covariance matrix is:

$$[\prod_i n_i^{k-1}(\sum n_i + 1)^{k-1}]/12^{\binom{k}{2}}.$$

The elements in the inverse of the variance-covariance matrix are: Corresponding to the variance of T_{ij} ,

$$12(\sum n_\alpha + 1 - n_i - n_j)/n_i n_j(\sum n_\alpha + 1).$$

Corresponding to a covariance with the T's having a common subscript, h, in the same position:

$$-12/n_h(\sum n_\alpha + 1).$$

Corresponding to a covariance with the T's having a common subscript, i, in different positions, for example T_{hi} and T_{ij} :

$$12/n_i(\sum n_\alpha + 1).$$

Finally, the element corresponding to $\sigma_{gh,ij}$ is zero.

The determinant of the variance-covariance matrix may be evaluated as follows: from the ith row of the determinant, factor $n_1 n_{1+i}/12$, $i = 1, \cdots, k - 1$; from the $([k - 1] + i)$-th row, factor $n_2 n_{2+i}/12$, $i = 1, \cdots, k - 2$; $\cdots$; from the last row, factor $n_{k-1}n_k/12$. The product of these factors is $\prod_i n_i^{k-1}/12^{\binom{k}{2}}$.

The entries in the determinant which is the other factor may be described somewhat crudely as follows.

The i-th diagonal block, $i = 1, \cdots, k - 1$, which contains the variances and covariances of the T_{ij}'s, fixed i and $j > i$, will be

$$\begin{bmatrix} n_i + n_{i+1} + 1 & n_{i+2} & \cdots & n_k \\ n_{i+1} & n_i + n_{i+2} + 1 & \cdots & n_k \\ \cdots & \cdots & \cdots & \cdots \\ n_{i+1} & n_{i+2} & \cdots & n_i + n_k + 1 \end{bmatrix}.$$

This is a $(k - i) \times (k - i)$ block.

The block consisting of the same rows and the first $(k - 1)$ columns is

$$\begin{bmatrix} \underbrace{\begin{matrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ & \cdots & \\ 0 & \cdots & 0 \end{matrix}}_{i - 2 \text{ columns}} & \begin{matrix} -n_1 \\ -n_1 \\ \cdots \\ -n_1 \end{matrix} & \underbrace{\begin{matrix} n_1 & 0 & 0 & \cdots & 0 \\ 0 & n_1 & 0 & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & n_1 \end{matrix}}_{k - i \text{ columns}} \end{bmatrix}.$$

The next block to the right consists of $k - 2$ columns and is:

$$\begin{bmatrix} \begin{matrix} 0 & \cdots & 0 \end{matrix} & -n_2 & \begin{matrix} n_2 & 0 & 0 & \cdots & 0 \end{matrix} \\ \begin{matrix} 0 & \cdots & 0 \end{matrix} & -n_2 & \begin{matrix} 0 & n_2 & 0 & \cdots & 0 \end{matrix} \\ \cdots & \cdots & \cdots \\ \underbrace{\begin{matrix} 0 & \cdots & 0 \end{matrix}}_{i-3 \text{ columns}} & -n_2 & \underbrace{\begin{matrix} 0 & 0 & & \cdots & n_2 \end{matrix}}_{k-i \text{ columns}} \end{bmatrix}.$$

The pattern is now clear.

The first block to the right of the i-th diagonal block consists of the next $(k - [i + 1])$ columns and is

$$\begin{bmatrix} -n_{i+2} & -n_{i+3} & \cdots & -n_k \\ n_{i+1} & 0 & \cdots & 0 \\ 0 & n_{i+1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & n_{i+1} \end{bmatrix}.$$

The block to the right of this is

$$\begin{bmatrix} 0 & 0 & \cdots & 0 \\ -n_{i+3} & -n_{i+4} & \cdots & -n_k \\ n_{i+2} & 0 & \cdots & 0 \\ 0 & n_{i+2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & n_{i+2} \end{bmatrix}.$$

Again, the pattern is clear.

At the next step in evaluating the determinant, from column 1 subtract columns $([k-1]+1)$ through $([k-1]+[k-2])$. To column 2, add column $([k - 1] + 1)$ and subtract columns $([k - 1] + [k - 2] + 1)$ through $([k - 1] + [k - 2] + [k - 3])$. In general, to column i, $i = 1, \cdots, k - 1$, add the $i - 1$ columns described as those containing the first column of the diagonal blocks previously set out, and subtract the sum of the next $(k - [i + 1])$ columns, that is, the columns including the $(i + 1)$-st diagonal block.

The result of the above operations is that the first diagonal block has $\sum n_i + 1$ as common diagonal element and zeros elsewhere. The other blocks containing the first $k - 1$ columns are of the form previously described but with $-n_1$ and n_1 replaced by $-(\sum n_i + 1)$ and $\sum n_i + 1$ respectively.

At the next and final step, leave rows $1, \cdots, k - 1$ unaltered.

Consider the next $k - 2$ rows which we now call the first set, the next $k - 3$ rows called the second set, and so on, the i-th set consisting of $(k - [i + 1])$ rows, beginning with row $\sum_{\alpha=1}^{i} (k - \alpha) + 1$. To the j-the row of the i-th set, add row i and subtract row $i + j$, these rows coming from the first $k - 1$ rows of the determinant.

The result of these operations is a determinant with $\sum n_i + 1$ in the first $k - 1$ principal diagonal positions, ones elsewhere in the principal diagonal, and zeros below the principal diagonal. Hence the determinant is as given.

That the inverse elements are correctly given may be checked by multiplying the matrix by the given inverse.

For the treatments versus control procedure, let X_i, $i = 0, 1, \cdots, k$ be random variables measuring some characteristic of a control and k treatments with n_i observations in the i-th sample.

Computation of the test criterion requires us to:

1. Rank jointly the X_0's and X_i's, i fixed, giving rank 1 to the least observation.
2. Add ranks for the variable with fewer observations, here assumed to be the check, to give T_i.
3. Compute the conjugate, T'_i.

Consider $(T_1, \cdots, T_k)$. Again, a recursion formula for finding the number of permutations which give rise to a specific value of $(T_1, \cdots, T_k)$ is given in [13].

The distribution of $(T_1, \cdots, T_k)$ has been shown to have the following parameters [13]:

$$\begin{aligned} E(T_i) &= \mu_i = n_0(n_0 + n_i + 1)/2, \\ E(T_i - \mu_i)^2 &= \sigma_i^2 = n_0 n_i(n_0 + n_i + 1)/12, \\ E(T_i - \mu_i)(T_j - \mu_j) &= \sigma_{ij} = n_0 n_i n_j/12, \\ \rho_{ij}^2 &= n_i n_j/(n_0 + n_i + 1)(n_0 + n_j + 1). \end{aligned}$$

It may also be shown that the determinant of the variance-covariance matrix is

$$\prod_0^k n_i(n_0[n_0 + 1])^{k-1}\left(\sum_0^k n_i + 1\right)/12.$$

The diagonal and off-diagonal elements of the inverse are, respectively,

$$12\left(\sum_{\alpha \neq i} n_\alpha + 1\right) \Big/ \left[n_0(n_0 + 1)n_i\left(\sum_0^k n_\alpha + 1\right)\right],$$

and

$$-12 \Big/ \left[n_0(n_0 + 1)\left(\sum_0^k n_i + 1\right)\right].$$

To evaluate the determinant of the variance-covariance matrix, factor $n_0 n_i/12$ from the i-th row of the determinant.
This gives:

$$\frac{n_0^{k-1} \prod_0^k n_i}{12^k} \begin{vmatrix} n_0 + n_1 + 1 & n_2 & \cdots & n_k \\ n_1 & n_0 + n_2 + 1 & \cdots & n_k \\ \cdots & \cdots & \cdots & \cdots \\ n_1 & n_2 & \cdots & n_0 + n_k + 1 \end{vmatrix}.$$

Next, subtract the k-th row from the i-th row, $i = 1, \cdots, k - 1$. We obtain

$$\frac{n_0^{k-1} \prod_0^k n_i}{12^k} \begin{vmatrix} n_0 + 1 & 0 & \cdots & -(n_0 + 1) \\ 0 & n_0 + 1 & \cdots & -(n_0 + 1) \\ \cdots & \cdots & \cdots & \cdots \\ n_1 & n_2 & \cdots & n_0 + n_k + 1 \end{vmatrix}.$$

Finally, obtain a new k-th column as the sum of all columns. It is then apparent that the given determinant is correct.

That the given inverse elements are correct is seen by multiplying the matrix by the stated inverse.

From the above information, it is possible to tabulate probabilities for the parent distribution and derived distributions of interest.

6. ACKNOWLEDGEMENT

I wish to thank Dr. Richard C. Lesser, Director, Cornell Computing Center, Cornell University, Ithaca, New York, for making machine time and programming help available for the random sampling.

A considerable portion of the work in this paper was done while I was a member of the Biometrics Unit, Department of Plant Breeding, Cornell University.

7. REFERENCES

[1] Duncan, D. B. [1955]. Multiple range and multiple F tests. *Biometrics 11*, 1–42.

[2] Dunnett, C. W. [1955]. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Stat. Assoc. 50*, 1096–1121.

[3] Fraser, D. A. S. [1956]. A vector form of the Wald-Wolfowitz-Hoeffding theorem. *Ann. Math. Stat. 27*, 540–43.

[4] Harter, H. L. [1957]. Error rates and sample sizes for range tests in multiple comparisons. *Biometrics 13*, 511–36.

[5] Harter, H. L. [1960]. Critical values for Duncan's new multiple range test. *Biometrics 16*, 671–85.

[6] Harter, H. L. [1961]. Note 161-Corrected error rates for Duncan's new multiple range test. *Biometrics 17*, 321–24.

[7] Keuls, M. [1952]. The use of the 'studentized range' in connection with an analy sis of variance. *Euphytica 1*, 112–22.

[8] Kolmogorov, A. [1941]. Confidence limits for an unknown distribution function. *Ann. Math. Stat. 12*, 461–63.

[9] Kruskal, W. H., and W. A. Wallis. [1952]. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc. 47*, 583–621.

[10] Mann, H. B., and D. R. Whitney. [1947]. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat. 18*, 50–60.

[11] Newman, D. [1939]. The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika 31*, 20–30.

[12] Pearson, E. S., and H. O. Hartley. [1954]. *Biometrika Tables for Statisticians*, Vol. I. Cambridge University Press.

[13] Pfanzagl, J. [1959]. Ein kombiniertes Test und Klassifikations—Problem. *Metrika 2*, 11–45.

[14] Smirnov, N. [1948]. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat. 19*, 279–81.

[15] Snedecor, G. W. [1948]. Answer to Query 60. *Biometrics 4*, 213–15.

[16] Steel, R. G. D. [1959]. A multiple comparison rank sum test: Treatments versus control. *Biometrics 15*, 560–72.

[17] Steel, R. G. D. [1960]. A rank sum test for comparing all pairs of treatments. *Technometrics 2*, 197–207.

[18] Steel, R. G. D. [1961]. Answer to Query 163. *Biometrics 17*, 326–28.

[19] Steel, R. G. D., and J. H. Torrie. [1960]. *Principles and procedures of statistics*, McGraw-Hill Book Company, Inc., New York.

[20] White, C. [1952]. The use of ranks in a test of significance for comparing two treatments. *Biometrics 8*, 33–41.

[21] Wilcoxon, F. [1945]. Individual comparisons by ranking methods. *Biometrics 1*, 80–3.

THE ESTIMATION OF REPEATABILITY AND HERITABILITY FROM RECORDS SUBJECT TO CULLING

R. N. Curnow[1]
Agricultural Research Council Unit of Statistics,
University of Aberdeen, Aberdeen, Scotland[2]

1. INTRODUCTION

In any animal breeding selection programme, estimates of repeatability and heritability are needed to choose between the various selection schemes available and also to ensure that the highest possible genetic gains are obtained from the chosen scheme. Estimates of repeatability and heritability are generally subject to large sampling errors. Therefore, the most efficient methods of estimation should be used even if they do involve rather lengthy computations. In this paper, the maximum likelihood estimation of repeatability and heritability from records subject to culling will be considered. The more usual regression estimators are often very inefficient compared with these maximum likelihood estimators.

Suppose that we wish to estimate the repeatability of lactation yield in a herd of dairy cattle. We shall assume that only first and second lactation yields are available and that, if there had been no culling (*i.e.*, if all the cows had had second records as well as first records), the first and second records would have been normally distributed over the herd with means μ_1 and μ_2 , variances σ_1^2 and σ_2^2 and covariance between the two records of the same cow σ_{12} . The first and second records of cow i will be written y_{i1} and y_{i2} respectively. The assumption of normality for the distributions will probably be a reasonably good approximation unless the herd can be split into groups so that any two cows in the same group are much more alike than two cows in different groups. These groups may, for example, be groups of daughters of the same sire or groups of cows according to the year in which they gave their first record or the month in which they calved. Methods are available for making allowances for such groupings, but they will be assumed absent in the rest of this paper. Very rarely will the culling

[1] The work on this paper was undertaken while the author was visiting the United States on a Harkness Fellowship of the Commonwealth Fund.

[2] Present address: Unit of Biometry, University of Reading.

in the herd be sufficiently intense or sufficiently highly correlated with future milk yielding capacity to affect seriously the normality of the distributions.

Repeatability is defined as the correlation between two different records of the same cow and is therefore

$$\rho = \sigma_{12}/\sigma_1\sigma_2 \ .$$

Since we are considering only first and second records, the question of whether the repeatability is the same for all pairs of records will not be discussed. The assumption is frequently made that $\sigma_1^2 = \sigma_2^2$ and, therefore, that repeatability is the same quantity as the regression coefficient of second records on first, $\beta_{21} = \sigma_{12}/\sigma_1^2$. The assumption is made, for example, in the formula given by Lush [1945] for comparing cows with a differing number of records and in the formula given by Lerner [1958] for the ratio of the heritability of the mean of n records to the heritability of a single record. It was also used by Henderson, Kempthorne, Searle and von Krosigk [1959] in their discussion of the disentanglement of environmental and genetic trends from records subject to culling. When $\sigma_1^2 \neq \sigma_2^2$, an estimate of β_{21} is needed for prediction purposes but for other purposes estimates of σ_1^2 and σ_2^2 may also be required. We shall assume in this paper that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, say, and therefore that $\beta_{21} = \rho$. A logarithmic transformation applied to the data may sometimes be useful in satisfying this assumption.

ρ is generally estimated by b, the sample regression coefficient of second records on first. There are two reasons for this. First, b is very simple to calculate and, second, it is an unbiassed estimator of ρ despite any culling that may have been based on the first records. However, when $\sigma_1^2 = \sigma_2^2$ and the first records of all cows are available, whether or not they have second records, b is not the maximum likelihood estimator of ρ. The variance of the second records about their regression on the first estimates $\sigma^2(1 - \rho^2)$ and the variance of all the first records estimates σ^2. These two estimators can be combined to give an estimator of ρ^2. Maximum likelihood makes use of this information as well as the information given by b.

We shall derive the maximum likelihood estimators of ρ and σ^2. We shall show that the efficiency of b as an estimator of ρ, relative to the maximum likelihood estimator, is fairly low for values of the various parameters that may well occur in practice. The bias of the maximum likelihood estimator is shown to be small. This suggests that the maximum likelihood estimator may be worth calculating. The computations involve only the solution of a cubic equation. A section is devoted to an approximate check of the assumption that $\sigma_1^2 = \sigma_2^2$.

Attention has been confined so far to the estimation of repeatability. The methods to be discussed could also be applied to the estimation of heritability from parent-offspring records. The assumption $\sigma_1^2 = \sigma_2^2$ means that, had there been no selection of parents, the variance of the parents and of the offspring would have been equal. In heritability studies, the maximum likelihood method makes use of information about animals that are not parents of animals which also have records. In milk yield studies, this would include information on dams that have only male calves.

2. THE MAXIMUM LIKELIHOOD ESTIMATION OF REPEATABILITY

All the cows have first lactation records but they do not all have second lactation records. We shall assume that the probability that a cow has a second record depends on its first record but not on any other character correlated with the second record. This rules out culling based on information about relatives or on characters such as percentage butter-fat. However, the effect of such culling on the estimates will often be very small and could probably be safely ignored.

Let N cows have a first record and $n \leq N$ cows have a second record. Numbering the cows with a second record $i = 1, 2, \cdots, n$ and the cows without a second record $i = n + 1, n + 2, \cdots, N$, the first records can be written

$$y_{i1} \qquad (i = 1, 2, \cdots, N)$$

and the second records

$$y_{i2} \qquad (i = 1, 2, \cdots, n).$$

The N first records are normally distributed with mean μ_1 and variance σ^2. Because the culling is based only on first records, the distribution of a second record y_{i2}, given the first y_{i1}, is independent of the distribution of the first record and is normal with mean $\mu_2 + \rho(y_{i1} - \mu_1)$ and variance $\sigma^2(1 - \rho^2)$. The likelihood of all the records is therefore

$$L = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{\sum_{i=1}^{N}(y_{i1} - \mu_1)^2}{2\sigma^2}\right\}$$
$$\times \frac{1}{[2\pi\sigma^2(1 - \rho^2)]^{n/2}} \exp\left\{-\frac{\sum_{i=1}^{n}[y_{i2} - \mu_2 - \rho(y_{i1} - \mu_1)]^2}{2\sigma^2(1 - \rho^2)}\right\}$$

and the log likelihood is, apart from a constant,

$$\ln L = -\frac{(n+N)}{2}\ln\sigma^2 - \frac{n}{2}\ln(1-\rho^2) - \frac{(1-\rho^2)\sum_{i=1}^{N}(y_{i1}-\mu_1)^2 + \sum_{i=1}^{n}[y_{i2}-\mu_2-\rho(y_{i1}-\mu_1)]^2}{2\sigma^2(1-\rho^2)}. \tag{2.1}$$

Kempthorne and von Krosigk (Henderson, Kempthorne, Searle and von Krosigk [1959]) suggested the use of maximum likelihood to estimate repeatability from records subject to culling. They derived the log likelihood when the cows were classified into groups and when the number of records for each cow was perfectly general. They derived very lengthy equations from which the maximum likelihood estimates of the parameters could be determined. In this paper, we are studying in greater detail the special case when there is no grouping of the cows and when there are at most only two records per cow.

(2.1) can be written

$$\begin{aligned}\ln L = -\frac{(n+N)}{2}\ln\sigma^2 - \frac{n}{2}\ln(1-\rho^2) - \frac{Ns^2}{2\sigma^2}\\ - \frac{n(\rho^2\Sigma_1^2+\Sigma_2^2-2\rho\Sigma_{12})}{2\sigma^2(1-\rho^2)}\\ - \frac{N(\bar{y}-\mu_1)^2}{2\sigma^2} - \frac{n[\bar{y}_2-\mu_2-\rho(\bar{y}_1-\mu_1)]^2}{2\sigma^2(1-\rho^2)}\end{aligned} \tag{2.2}$$

where

$$s^2 = \frac{\sum_{i=1}^{N}(y_{i1}-\bar{y})^2}{N}, \qquad \Sigma_1^2 = \frac{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2}{n},$$

$$\Sigma_2^2 = \frac{\sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2}{n}, \qquad \Sigma_{12} = \frac{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2)}{n},$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_{i1}}{N}, \qquad \bar{y}_1 = \frac{\sum_{i=1}^{n} y_{i1}}{n} \quad \text{and} \quad \bar{y}_2 = \frac{\sum_{i=1}^{n} y_{i2}}{n}.$$

The maximum likelihood estimators of μ_1 and μ_2 are clearly given by

$$\hat{\mu}_1 = \bar{y} \quad \text{and} \quad \hat{\mu}_2 = \bar{y}_2 - \hat{\rho}(\bar{y}_1 - \mu_1).$$

Also

$$\begin{aligned}\frac{\partial(\ln L)}{\partial(\sigma^2)} = -\frac{(n+N)}{2\sigma^2} + \frac{Ns^2}{2\sigma^4} + \frac{n(\rho^2\Sigma_1^2+\Sigma_2^2-2\rho\Sigma_{12})}{2\sigma^4(1-\rho^2)}\\ + \frac{N(\bar{y}-\mu_1)^2}{2\sigma^4} + \frac{n[\bar{y}_2-\mu_2-\rho(\bar{y}_1-\mu_1)]^2}{2\sigma^4(1-\rho^2)}\end{aligned} \tag{2.3}$$

and

$$\frac{\partial(\ln L)}{\partial\rho} = \frac{n\rho}{1-\rho^2} - \frac{n[\rho(\Sigma_1^2 + \Sigma_2^2) - (1+\rho^2)\Sigma_{12}]}{\sigma^2(1-\rho^2)^2} + \frac{n(\bar{y}_1 - \mu_1)[\bar{y}_2 - \mu_2 - \rho(\bar{y}_1 - \mu_1)]}{\sigma^2(1-\rho^2)} - \frac{n\rho[\bar{y}_2 - \mu_2 - \rho(\bar{y}_1 - \mu_1)]^2}{\sigma^2(1-\rho^2)^2}. \tag{2.4}$$

By setting (2.3) and (2.4) equal to zero and substituting the maximum likelihood estimators of μ_1 and μ_2, $\hat{\rho}$ satisfies the cubic

$$(Ns^2 - n\Sigma_1^2)\hat{\rho}^3 - (N-n)\Sigma_{12}\hat{\rho}^2 + [(n+N)\Sigma_1^2 - N(s^2 - \Sigma_2^2)]\hat{\rho} - (n+N)\Sigma_{12} = 0. \tag{2.5}$$

This equation will have one and only one root between -1 and $+1$ of the same sign as the simple estimator $b = \Sigma_{12}/\Sigma_1^2$. There may be two other roots between -1 and $+1$ of opposite sign to b. The formula for $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{\hat{\rho}(\Sigma_1^2 + \Sigma_2^2) - (1+\hat{\rho}^2)\Sigma_{12}}{\hat{\rho}(1-\hat{\rho}^2)}. \tag{2.6}$$

When there is no culling, $s^2 = \Sigma_1^2$ and

$$\hat{\rho} = \frac{2\Sigma_{12}}{\Sigma_1^2 + \Sigma_2^2}.$$

When $\sigma_1^2 = \sigma_2^2$, the maximum likelihood estimator of the correlation coefficient has the arithmetic rather than the geometric mean of the two sample variances in the denominator.

To obtain the asymptotic variance of $\hat{\rho}$, we need the expected values of the second-order partial derivatives of ln L with respect to μ_1, μ_2, ρ and σ^2. These can be derived from a knowledge of the expected values of s^2, Σ_2^2 and Σ_{12}. Σ_1^2 depends on the method of culling and so will be taken as fixed. For reasons to be given later, we shall calculate the expected values when $\sigma_1^2 \neq \sigma_2^2$. Clearly, $E(s^2) = (N - 1/N)\sigma_1^2$. By writing

$$y_{i2} = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(y_{i1} - \mu_1) + \sigma_2\sqrt{1-\rho^2}\, e_{i2.1},$$

so that $e_{i2\ 1}$ has a standard normal distribution independent of $y_{.1}$,

$$\Sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\rho\sigma_2}{\sigma_1}(y_{i1} - \bar{y}_1) + \sigma_2\sqrt{1-\rho^2}\,(e_{i2.1} - \bar{e}_{.2.1})\right]^2,$$

and, therefore, for given Σ_1^2,

$$E(\Sigma_2^2) = \frac{\rho^2\sigma_2^2}{\sigma_1^2}\,\Sigma_1^2 + \frac{(n-1)}{n}\,\sigma_2^2(1-\rho^2).$$

Similarly,

$$E(\Sigma_{12}) = \frac{\rho\sigma_2}{\sigma_1}\,\Sigma_1^2\ .$$

Also, with $\sigma_1^2 = \sigma_2^2$, $E(\bar{y} - \mu_1)^2 = \sigma^2/N$, $E[\bar{y}_2 - \mu_2 - \rho(\bar{y}_1 - \mu_1)]^2 = \sigma^2(1-\rho^2)/n$ and $E\{(\bar{y}_1 - \mu_1)[\bar{y}_2 - \mu_2 - \rho(\bar{y}_1 - \mu_1)]\} = 0$. After considerable algebra the asymptotic variance—covariance matrix for $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}^2$ and $\hat{\rho}$ is found to be **V**, where

$$\mathbf{V}^{-1} = \begin{bmatrix} \dfrac{N}{\sigma^2} + \dfrac{n\rho^2}{\sigma^2(1-\rho^2)} & -\dfrac{n\rho}{\sigma^2(1-\rho^2)} & 0 & -\dfrac{n\rho(\bar{y}_1 - \mu_1)}{\sigma^2(1-\rho^2)} \\ -\dfrac{n\rho}{\sigma^2(1-\rho^2)} & \dfrac{n}{\sigma^2(1-\rho^2)} & 0 & \dfrac{n(\bar{y}_1 - \mu_1)}{\sigma^2(1-\rho^2)} \\ 0 & 0 & \dfrac{n+N}{2\sigma^4} & -\dfrac{n\rho}{\sigma^2(1-\rho^2)} \\ -\dfrac{n\rho(\bar{y}_1 - \mu_1)}{\sigma^2(1-\rho^2)} & \dfrac{n(\bar{y}_1 - \mu_1)}{\sigma^2(1-\rho^2)} & -\dfrac{n\rho}{\sigma^2(1-\rho^2)} & \theta \end{bmatrix}$$

and

$$\theta = \frac{n}{\sigma^2(1-\rho^2)^2}\,[(1-\rho^2)\Sigma_1^2 + 2\rho^2\sigma^2] + \frac{n(\bar{y}_1 - \mu_1)^2}{\sigma^2(1-\rho^2)}.$$

The asymptotic variance of $\hat{\rho}$ is

$$V(\hat{\rho}) = \frac{(N+n)(1-\rho^2)^2}{n[(N+n)(1-\rho^2)\Sigma_1^2/\sigma^2 + 2N\rho^2]}.$$

Writing $n/N = S$, so that $(1 - S)$ is the culling intensity, and $\Sigma_1^2/\sigma^2 = 1/c$, where c measures the effect of the culling on the variance of the first records and is therefore a possible measure of the efficiency of the culling,

$$V(\hat{\rho}) = \frac{1}{n}\,\frac{(1+S)(1-\rho^2)^2}{\left[\dfrac{(1+S)(1-\rho^2)}{c} + 2\rho^2\right]}. \tag{2.7}$$

We shall now derive the approximate bias in $\hat{\rho}$ as an estimator of ρ. By substituting $\hat{\rho} = \rho + \delta$ in (2.5) and ignoring terms in δ^2 and δ^3,

$$\delta \doteq -\frac{(Ns^2 - n\Sigma_1^2)\rho^3 - (N-n)\Sigma_{12}\rho^2 + [(n+N)\Sigma_1^2 - N(s^2 - \Sigma_2^2)]\rho - (n+N)\Sigma_{12}}{3(Ns^2 - n\Sigma_1{}^2)\rho^2 - 2(N-n)\Sigma_{12}\rho + [(n+N)\Sigma_1{}^2 - N(s^2 - \Sigma_2^2\ )]}. \tag{2.8}$$

Approximating $E(\delta)$ by taking expected values separately in the numerator and denominator,

$$E(\delta) \doteq -\frac{\rho(1-\rho^2)\sigma^2(n-N)}{\Sigma_1^2 n(n+N)(1-\rho^2)+\sigma^2[(n-N)+\rho^2(2nN+N-3n)]}.$$

To order $1/n$,

$$E(\delta) = \frac{\rho(1-\rho^2)(1-S)}{n\left[\frac{(1+S)(1-\rho^2)}{c}+2\rho^2\right]}. \tag{2.9}$$

Therefore, from (2.7), the approximate ratio of the bias to the standard deviation is

$$\frac{E(\delta)}{\sqrt{V(\hat{\rho})}} \doteq \frac{\rho(1-S)}{\sqrt{n}\,\sqrt{1+S}\left\{\frac{1+S}{c}(1-\rho^2)+2\rho^2\right\}^{1/2}}$$

$$\leq \frac{\rho(1-S)}{\sqrt{n}\,\sqrt{1+S}}\,\mathrm{Max}\left[\sqrt{\frac{c}{1+S}},\,1/\sqrt{2}\right].$$

If n is reasonably large, the bias in $\hat{\rho}$ is unlikely to be serious. The bias will become relatively more important when estimates of ρ from different sources are pooled.

3. ESTIMATION OF σ_1^2 AND σ_2^2 WHEN $\sigma_1^2 \neq \sigma_2^2$

In the previous section we derived the expected values of s^2, Σ_2^2 and Σ_{12} for given Σ_1^2 when $\sigma_1^2 \neq \sigma_2^2$. They were

$$E(s^2) = \frac{N-1}{N}\sigma_1^2,$$

$$E(\Sigma_2^2) = \frac{\rho^2\sigma_2^2}{\sigma_1^2}\Sigma_1^2 + \frac{(n-1)}{n}\sigma_2^2(1-\rho^2),$$

and

$$E(\Sigma_{12}) = \frac{\rho\sigma_2}{\sigma_1}\Sigma_1^2.$$

Apart from the usual slight differences in the multipliers, important only for small n and N, the following estimators of σ_1^2, $\beta = \rho\sigma_2/\sigma_1$ and $\sigma_2^2(1-\rho^2)$ are the maximum likelihood estimators when $\sigma_1^2 \neq \sigma_2^2$,

$$\hat{\sigma}_1^2 = Ns^2/(N-1),$$

$$\hat{\beta}_{21} = b = \Sigma_{12}/\Sigma_1^2,$$

and

$$\widehat{\sigma_2^2(1-\rho^2)} = \frac{n}{n-2}\,[\Sigma_2^2 - \Sigma_{12}^2/\Sigma_1^2].$$

From the expected values of s^2, Σ_2^2 and Σ_{12} given in the previous section these estimators are all unbiassed. The maximum likelihood estimator of σ_2^2/σ_1^2 is

$$\widehat{\left(\frac{\sigma_2^2}{\sigma_1^2}\right)} = \frac{\Sigma_{12}^2}{\Sigma_1^4} + \frac{\Sigma_2^2 - \Sigma_{12}^2/\Sigma_1^2}{s^2}. \tag{3.1}$$

No exact method is available for constructing a confidence interval for σ_2^2/σ_1^2 except when $n = N$, *i.e.*, when there is no culling (see Curnow [1957] for references and for a method to be used when the data are grouped). The asymptotic variance of $\widehat{(\sigma_2^2/\sigma_1^2)}$ could be derived and used to provide an approximate confidence interval. This would give some indication of the importance of the assumption that $\sigma_1^2 = \sigma_2^2$. A simpler, but much less sensitive, test of whether $\sigma_2^2 > \sigma_1^2$ could be based on the fact that the quantity

$$\frac{(N-1)n\sigma_1^2}{N(n-2)\sigma_2^2(1-\rho^2)} \times \frac{\Sigma_2^2 - \Sigma_{12}^2/\Sigma_1^2}{s^2},$$

which compares a χ^2-value from the variation of the second records about their regression on the first with a χ^2-value from the variation of all the first records, has an F-distribution with $n - 2$ and $N - 1$ degrees of freedom.

$$F = \frac{(N-1)n}{N(n-2)}\,\frac{\Sigma_2^2 - \Sigma_{12}^2/\Sigma_1^2}{s^2}$$

significantly greater than $F = 1$ would suggest that $\sigma_2^2(1-\rho^2) > \sigma_1^2$ and, therefore, that $\sigma_2^2 > \sigma_1^2$.

4. THE EFFICIENCY OF THE SIMPLE REGRESSION ESTIMATOR OF REPEATABILITY

The statistic $b = \Sigma_{12}/\Sigma_1^2$ is very easy to calculate and is the one generally used to estimate repeatability. It is the only possible estimator if first records are available only for those cows also having second records. Providing that culling is based solely on the first lactation yields, b is always an unbiassed estimator of $\beta_{21} = \rho\sigma_2/\sigma_1$, but an unbiassed estimator of ρ only when $\sigma_1^2 = \sigma_2^2$. In this paper we are assuming $\sigma_1^2 = \sigma_2^2$. The variance of b is

$$V(b) = \sigma^2/n\Sigma_1^2 = c/n. \tag{4.1}$$

From (2.7), the asymptotic efficiency of b relative to the maximum likelihood estimator, $\hat{\rho}$, is therefore

$$Eff. = \frac{(1 + S)(1 - \rho^2)^2}{(1 + S)(1 - \rho^2) + 2c\rho^2}.$$

The values of this efficiency are shown in Table 1 for various values of S, ρ and c.

The following considerations suggest that c is unlikely to be greater than $c = 2$. Let the N cows be reduced to LN by accidental factors uncorrelated with the level of yield. Then a proportion $P = n/LN = S/L$ can be selected on the basis of first lactation yield. Assume that the selection is of the proportion S/L of the herd having the highest yields and that n and LN are sufficiently large that the effect of the selection

TABLE 1

THE ASYMPTOTIC EFFICIENCY OF THE SIMPLE REGRESSION ESTIMATOR OF REPEATABILITY RELATIVE TO THE MAXIMUM LIKELIHOOD ESTIMATOR

Overall Selection Intensity ($S = n/N$)	Repeatability (ρ)	$c = \sigma^2/\Sigma_1^2$				
		1/2	1	4/3	2	4
1/4	0	1.00	1.00	1.00	1.00	1.00
	1/4	0.95	0.90	0.88	0.82	0.70
	1/2	0.79	0.65	0.58	0.48	0.32
	3/4	0.49	0.33	0.27	0.20	0.11
	1	0	0	0	0	0
1/2	0	1.00	1.00	1.00	1.00	1.00
	1/4	0.96	0.92	0.89	0.85	0.74
	1/2	0.82	0.69	0.63	0.53	0.36
	3/4	0.54	0.37	0.30	0.23	0.13
	1	0	0	0	0	0
3/4	0	1.00	1.00	1.00	1.00	1.00
	1/4	0.96	0.93	0.91	0.87	0.77
	1/2	0.84	0.72	0.66	0.57	0.40
	3/4	0.58	0.40	0.34	0.25	0.15
	1	0	0	0	0	0
1	0	1.00	1.00	1.00	1.00	1.00
	1/4	0.97	0.94	0.92	0.88	0.79
	1/2	0.86	0.75	0.69	0.60	0.43
	3/4	0.61	0.44	0.37	0.28	0.16
	1	0	0	0	0	0

can be approximated by the truncation of the top P proportion of an infinite normal population with mean μ_1 and variance σ^2. The variance of the selected first records will then be

$$\Sigma_1^2 = \sigma^2[1 - \nu(\nu - T)],$$

where $\nu = Z/P$ and Z and T are the ordinate and abscissa at the point above which lies a proportion P of the population (Finney [1957] derives this formula and provides a table of values of $\nu' = \nu(\nu - T)$ for various values of P). The division of the overall selection intensity, S, between L and P is never complete and is certainly never known exactly. However, P is generally considerably larger than S, *i.e.*, there is a large amount of random culling and not so much actual selection. The value of c is

$$c = \frac{\sigma^2}{\Sigma_1^2} = 1/[1 - \nu(\nu - T)].$$

For various P, it takes the following values:

$P =$ 0.95	0.9	0.8	0.7	0.6	0.5	0.1,
$c =$ 1.24	1.40	1.72	2.03	2.37	2.75	5.91.

In practice, P is unlikely to be less than $P = 0.7$ and therefore c is unlikely to exceed $c = 2$. Table 1 does include $c = 4$ to illustrate the effect of intense selection based on yield and $c = \frac{1}{2}$ to illustrate the effect of a selection scheme that results in an increased rather than a decreased variance. The efficiency is the same for ρ as for $-\rho$ and so ρ is shown as positive in the table. In practice, ρ is very unlikely to be negative.

When $\rho = 0$, the efficiency of the simple regression estimator is 1. When $\rho = 1$, it is 0. For intermediate values of ρ, the efficiency increases rather slowly with S for fixed c, but decreases fairly rapidly with increasing c. For all values of c and S, the efficiency is very low for high values of ρ. As an example of the efficiency likely to occur in the estimation of the repeatability of lactation records, the efficiency is 0.66 when $c = \frac{4}{3}$, $\rho = \frac{1}{2}$ and $S = \frac{3}{4}$. In this case, the maximum likelihood estimator would certainly be worth calculating.

5. THE ESTIMATION OF REPEATABILITY BY AN ANALYSIS OF VARIANCE

Sometimes the effect of culling based on the first records is ignored and ρ estimated from a least squares analysis of variance of the $N + n$ first and second records. Table 2 shows this analysis of variance. The "cows and periods" sum of squares has been split into both "cows adjusting for periods and periods ignoring cows" and "cows ignoring periods and periods adjusting for cows". By a period difference is meant an average difference between first and second lactation records. The expected values of four important mean squares are shown. Δ_1, Δ_2 and Δ_3 are defined at the foot of the table and vanish only if the culling is random with respect to the first records. Two cases need to be distinguished. In the first we assume $\mu_1 = \mu_2$ and in the second we do not. We shall use the reference numbers at the right of the table. If $\mu_1 = \mu_2$, mean square 2 and a mean square obtained by pooling sums of squares 3 and 4 are used to estimate σ^2 and ρ. This is exactly equivalent to an analysis of a one way classification with unequal sub-class numbers (Snedecor, [1956], §10.16). If $\mu_1 \neq \mu_2$, mean squares 1 and 4 are used to estimate σ^2 and ρ. There seems to be little justification for using mean square 2 with mean square 4. One or other is biassed or inefficient according as $\mu_1 \neq \mu_2$ or $\mu_1 = \mu_2$.

The bias in any of the above methods is difficult to determine without some knowledge of the effect of culling on the expected values of $\sum_1^2 - [(n-1)/n]\sigma^2$, $\bar{y}_1 - \mu_1$, $(\bar{y}_1 - \mu_1)^2 - \sigma^2/n$ and $(\bar{y}_1 - \mu_1)\bar{y} - \sigma^2/N$. However, these unknown biasses are clearly undesirable. Since the method of estimation described in this paper is the maximum likelihood method whether or not the culling is random with respect to first records, it is to be preferred. If culling is random with respect to first records, c can be substituted as $c = 1$ in the formulae for the asymptotic bias and variance of the maximum likelihood estimator and the asymptotic efficiency of the regression method.

Wadell [1961] has given a method for estimating ρ when the culling is equivalent to truncation of the first lactation yields. His estimate of ρ is a function of $\bar{y}_1$, $\bar{y}_2$, s^2, Σ_1^2 and the mean squares 2, 3, and 4 of Table 2. The estimate has a negligible bias for large values of n but its variance is not given and therefore its efficiency is unknown.

6. A NUMERICAL EXAMPLE

Dr. A. E. Freeman of the Department of Animal Husbandry, Iowa State University has kindly made available to me some Complete Herd Improvement Registry Records of an Iowa Board of Control herd of Holstein-Friesian cattle. The records included some first and second lactation yields made in various years but expressed as deviations from the herd average for each particular year. The cows were milked twice daily and their yields expressed as mature equivalent 305-day lactation yields.

The values of the various relevant statistics for a sample of the records were:

TABLE 2

Analysis of Variance of First and Second Records

Source	d.f	Sums of Squares	Expected Values of Mean Squares	Ref. Nos.
Cows adj. Periods	$N - 1$	$\frac{n}{2}(\Sigma_2^2 + 2\Sigma_{12} - \Sigma_1^2) + Ns^2$	$\sigma^2(1 - \rho) + \frac{n + N - 2}{N - 1}\rho\sigma^2$ $+ \frac{n}{2(N - 1)}(\rho^2 + 2\rho - 1)\Delta_1$	1
Periods ign. Cows	1	$\frac{nN}{N + n}(\bar{y} - \bar{y}_2)^2$		
Cows and Periods	N	$\frac{n}{2}(\Sigma_2^2 + 2\Sigma_{12} - \Sigma_1^2) + Ns^2$ $+ \frac{nN}{N + n}(\bar{y} - \bar{y}_2)^2$		
Cows ign. Periods	$N - 1$	$\frac{n}{2}(\Sigma_2^2 + 2\Sigma_{12} - \Sigma_1^2) + Ns^2$ $+ \frac{nN}{N + n}(\bar{y} - \bar{y}_2)^2 - \frac{n}{2}(\bar{y}_1 - \bar{y}_2)^2$	$\sigma^2(1 - \rho)$ $+ \frac{(n + N)^2 - (N + 3n)}{(N - 1)(N + n)}\rho\sigma^2$ $+ \frac{n(N - n)}{2(N - 1)(N + n)}(\mu_2 - \mu_1)^2$ $+ \frac{n}{N - 1}\Delta_2$	2
Periods adj. Cows	1	$\frac{n}{2}(\bar{y}_1 - \bar{y}_2)^2$	$\sigma^2(1 - \rho) + \frac{n}{2}(\mu_2 - \mu_1)^2 + n(1 - \rho)\Delta_3$	3
Error	$n - 1$	$\frac{n}{2}(\Sigma_2^2 - 2\Sigma_{12} + \Sigma_1^2)$	$\sigma^2(1 - \rho) + \frac{n(1 - \rho)^2}{2(n - 1)}\Delta_1$	4
TOTAL	$N + n - 1$	$n\Sigma_2^2 + Ns^2 + \frac{Nn}{N + n}(\bar{y} - \bar{y}_2)^2$		
Adj. for mean	1	$\frac{(N\bar{y} + n\bar{y}_2)^2}{N + n}$		

$$\Delta_1 = \Sigma_1^2 - \frac{n - 1}{n}\sigma^2,$$

$$\Delta_2 = \frac{1}{2}(\rho^2 + 2\rho - 1)\Delta_1 + \frac{1}{2}\left(\frac{N - n}{N + n}\rho^2 + 2\rho - 1\right)(D^2 - \sigma^2/n)$$
$$+ \left(\frac{N - n}{N + n}\rho\mu_2 + \mu_2 - \mu_1 + \rho\mu_1\right)D - \frac{2N\rho}{N + n}(D\bar{y} - \sigma^2/N)$$

$$\Delta_3 = \tfrac{1}{2}(1 - \rho)(D^2 - \sigma^2/n) - (\mu_2 - \mu_1)D \quad \text{and} \quad D = \bar{y}_1 - \mu_1 .$$

$$N = 220, \qquad n = 150, \qquad S = 0.682;$$
$$\bar{y} = -1.118, \qquad \bar{y}_1 = 1.147, \qquad \bar{y}_2 = -0.207;$$
$$s^2 = 286.55, \qquad \Sigma_1^2 = 270.51, \qquad \Sigma_{12} = 125.85 \quad \text{and} \quad \Sigma_2^2 = 354.98.$$

The maximum likelihood estimator of σ_2^2/σ_1^2 (3.1) is $\widehat{(\sigma_2^2/\sigma_1^2)} = 1.25$. This value is almost certainly not significantly greater than $\sigma_2^2/\sigma_1^2 = 1$. However, it is sufficiently large to raise some doubts about the assumption that $\sigma_2^2 = \sigma_1^2$. The estimate of $\sigma_2^2(1 - \rho^2)$ is greater, but not significantly greater, than the estimate of σ_1^2. However, for illustrative purposes we shall assume $\sigma_2^2 = \sigma_1^2$.

Equation (2.5) for $\hat{\rho}$ is

$$\hat{\rho}^3 - 0.3922\hat{\rho}^2 + 5.1257\hat{\rho} - 2.0728 = 0.$$

The simple regression estimator of ρ is

$$b = \Sigma_{12}/\Sigma_1^2 = 0.465.$$

Three cycles of an iteration based on formulae (2.8) and with $\hat{\rho}_0 = b$ as the initial solution show that the value of $\hat{\rho}$ is, to three decimal places,

$$\hat{\rho} = 0.404.$$

This is the only real root of the maximum likelihood equation for $\hat{\rho}$. The two estimators together with their estimated standard errors [(2.7) and (4.1)] are therefore:

$$b = 0.465 \pm 0.084$$

and

$$\hat{\rho} = 0.404 \pm 0.069.$$

The latter standard error is an asymptotic standard error. The estimated efficiency of the regression estimator is only 67 percent and so $\hat{\rho}$ is probably well worth calculating. The estimated asymptotic bias (2.9) is very small,

$$\widehat{E(\delta)} = 0.0004.$$

The maximum likelihood estimator of σ^2 (2.6) is $\hat{\sigma}^2 = 314.46$. This agrees well with the value $s^2 = 286.55$ used above in estimating $E(\delta)$ and the standard errors of b and $\hat{\rho}$.

The estimated value of c is

$$\hat{c} = s^2/\Sigma_1^2 = 1.059.$$

This value of c suggests that there has been very little selection based on first lactation yield in this particular sample of records. The standardized selection differential

$$(\bar{y}_1 - \bar{y})/s = 0.134$$

suggests that the effective selection intensity P was near 0.93.

7. SUMMARY

The maximum likelihood estimation of repeatability from first and second lactation records of a herd subject to culling is discussed. The

method is applicable only when it can be assumed that, had there been no culling, the variances of the first and the second lactation records would have been equal, and when all the first records are available whether or not the cow has a second record. The efficiency of the more usual estimate of repeatability, based on the regression of second records on first records, is shown to be low when the repeatability is high. In many cases the calculation of the maximum likelihood estimate would seem to be well worthwhile. The use of the method in estimating heritability is mentioned. An illustrative example is given.

ACKNOWLEDGEMENT

I am grateful to Professor O. Kempthorne for helpful discussions during the preparation of this paper.

REFERENCES

Curnow, R. N. [1957]. Heterogeneous error variances in split-plot experiments. *Biometrika 44*, 378–83.

Finney, D. J. [1957]. The consequences of selection for a variate subject to errors of measurement. *Revue de l'Institut International de Statistique 24*, 22–9.

Henderson, C. R., Kempthorne, O., Searle, S. R. and von Krosigk, C. M. [1959]. The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Lerner, I. M. [1958]. *The Genetic Basis of Selection.* John Wiley and Sons, Inc., New York.

Lush, J. L. [1945]. *Animal Breeding Plans.* 3rd Ed., Iowa State College Press, Ames, Iowa.

Snedecor, G. W. [1956]. *Statistical Methods.* 5th Ed., Iowa State College Press, Ames, Iowa.

Wadell, L. H. [1961]. Selection bias in intraclass correlation repeatability estimates. Unpublished manuscript.

THREE CLASSES OF UNIVARIATE DISCRETE DISTRIBUTIONS

C. G. Khatri and I. R. Patel[1]
M. S. University of Baroda, Baroda, India.

1. INTRODUCTION

Families of descrete distributions have been developed and studied by many authors, including, Neyman [1939], Feller [1943], Skellam [1952], Beall and Rescia [1953] and Gurland [1957, 1958]. These families are of three types:

$$\text{Type } A: \quad g_A(z) = \exp\{h(z)\},$$

$$\text{Type } B: \quad g_B(z) = \{h(z)\}^n,$$

$$\text{Type } C: \quad g_C(z) = c \log\{h(z)\},$$

where $g(z)$ represents a probability generating function (p.g.f.) and $h(z)$ is a p.g.f., except possibly for additive and multiplicative constants. The aim of this paper is to set up formulae for certain statistics for these types. It is hoped that these will be of use to reseach workers in practical fields, who will be formulating compound and generalised distributions of these types by using specific forms of $h(z)$.

2. RECURRENCE RELATIONS FOR PROBABILITIES

Notations: Let the r-th derivation of $f(z)$ be denoted as $f^{(r)}(z)$. Also let

$$P_r = \{g^{(r)}(z)/r!\}\,|_{z=0}\,, \tag{1}$$

and

$$\pi_r = \{h^{(r)}(z)/r!\}\,|_{z=0}\,. \tag{2}$$

By the definition of $g(z)$, it is clear that P_r denotes the probability of the r-th count in $g(z)$ and π_r , the probability of the r-th count, excepting possibly for additive and multiplicative constants.

Type A: Here

$$g_A(z) = \exp\{h(z)\}.$$

[1] Present address: Statistical Officer (I), Rajkot, India.

Successive differentiation leads to

$$g_A^{(r)}(z) = \sum_{k=1}^{r} \binom{r-1}{k-1} h^{(k)}(z) g_A^{(r-k)}(z). \tag{3}$$

Hence, on letting $z = 0$, we have

$$P_r = \sum_{k=1}^{r} k\pi_k P_{r-k}/r \quad \text{with} \quad P_0 = \exp(\pi_0). \tag{4}$$

Type B: Here

$$g_B(z) = \{h(z)\}^n.$$

Successive differentiation of $h(z)g_B^{(1)}(z) = ng_B(z)h^{(1)}(z)$ leads to

$$\sum_{k=1}^{r} \binom{r-1}{k-1} h^{(k-1)}(z) g_B^{(r-k+1)}(z) = n \sum_{k-1}^{r} \binom{r-1}{k-1} h^{(k)}(z) g_B^{(r-k)}(z). \tag{5}$$

Hence, on letting $z = 0$, we have

$$P_r = \sum_{k=1}^{r} (nk - r + k)\pi_k P_{r-k}/r\pi_0 \quad \text{with} \quad P_0 = \pi_0^n . \tag{6}$$

Type C: Here

$$g_C(z) = c \log \{h(z)\} \quad \text{where} \quad c = \log \{h(1)\}^{-1}.$$

Successive differentiation of $h(z)g_C^{(1)}(z) = ch^{(1)}(z)$ leads to

$$h(z)g_C^{(r)}(z) = ch^{(r)}(z) - \sum_{k=1}^{r-1} \binom{r-1}{k} h^{(k)}(z) g_C^{(r-k)}(z). \tag{7}$$

Hence, on letting $z = 0$, we have

$$P_r = \{rc\pi_r - \sum_{k=1}^{r-1} (r-k)\pi_k P_{r-k}\}/r\pi_0 \quad \text{with} \quad P_0 = c \log \pi_0 . \tag{8}$$

3. FACTORIAL CUMULANTS

Notations: Let

$$\mu'_{[r]} = \{h^{(r)}(z)\} \big|_{z=1} , \tag{9}$$

$$M'_{[r]} = \{g^{(r)}(z)\} \big|_{z=1} , \tag{10}$$

and

$$K_{[r]} = \{(d/dz)^r \log g(z)\} \big|_{z=1} . \tag{11}$$

From the definition of factorial cumulants, it is clear that $\mu'_{[r]}$ is the r-th factorial moment of $h(z)$ if $h(z)$ is a p.g.f. and $M'_{[r]}$ and $K_{[r]}$ are respectively r-th factorial moment and r-th factorial cumulant of $g(z)$.

Type A: Here $\log \{g_A(z)\} = h(z)$ and hence,

$$K_{[r]} = \mu'_{[r]} . \tag{12}$$

Type B: Here

$$\phi(z) = \log g_B(z) = n \log h(z).$$

Hence, on using (11), it is clear that r-th factorial cumulant of $g_B(z)$ is n times the r-th factorial cumulant of $h(z)$ if $h(z)$ is a p.g.f. Also on using (7) (with necessary modifications), (9) and (11), we have

$$K_{[r]} = n\mu'_{[r]} - \sum_{k=1}^{r-1} \binom{r-1}{k} \mu'_{[k]} K_{[r-k]} . \tag{13}$$

Type C: Here, it is easy to give a recurrence relation for factorial moments $M'_{[r]}$ rather than factorial cumulants. By using (7), (9) and (10), this relation can be shown to be

$$M'_{[r]} = \left\{ c\mu'_{[r]} - \sum_{k=1}^{r-1} \binom{r-1}{k} \mu'_{[k]} M'_{[r-k]} \right\} / \mu'_0 , \tag{14}$$

where $\mu'_0 = h(1)$. The factorial cumulants can be obtained from (14) by using the relations

$$K_{[1]} = M'_{[1]} , \qquad K_{[2]} = M'_{[2]} - M'^2_{[1]} ,$$
$$K_{[3]} = M'_{[3]} - 3M'_{[2]}M'_{[1]} + 2M'^3_1 , \qquad \text{etc.}$$

4. SPECIAL CASES

Notation: If a variate has either a Binomial or a Negative Binomial law as a special case, we say that it has a general binomial law.

Type A: From the form of the Poisson-Binomial, Negative-Binomial, generalised Polya Aeppli, Beall and Rescia [1953] and Neyman's Types A, B, and C, it is clear that they belong to this form. It is to be noted, that, if in the classical problem of egg masses and larvae, the egg masses have a Poisson distribution with p.g.f. $\exp [\lambda(z - 1)]$ and the larvae within an egg mass, a distribution with p.g.f. $w(z)$, then the distribution of the larvae over the whole field is $\exp [\lambda\{w(z) - 1\}]$ which is of Type A with $h(z) = \lambda\{w(z) - 1\}$.

Some important distributions with their recurrence relations are:

(i) Poisson-Hypergeometric distribution [1958]: Here

$$w(z) = \sum_{r=0}^{\infty} \binom{k}{r} \alpha_{(r)} m^r (z - 1)^r / (\alpha + \beta)_{(r)} \tag{15}$$

where

$$\alpha_{(r)} = \alpha(\alpha + 1)(\alpha + 2) \cdots (\alpha + r - 1), \qquad \binom{k}{r} = k^{(r)}/r!,$$

$k^{(r)} = k(k-1)\cdots(k-r+1)$, $\alpha_{(0)} = 1$, $k^{(0)} = 1$ and α, β, m and k are such that $w^{(r)}(0)/r! = \pi_r$ is positive. The above distribution was first given by Gurland [1958]. The recurrence relation in probabilities is

$$P_r = \lambda \sum_{s=1}^{r} s\pi_s P_{r-s}/r \tag{16}$$

where $P_0 = \exp\{\lambda(\pi_0 - 1)\}$, and

$$\pi_r = \frac{\alpha + \beta + r - 2 - m(\alpha - k + 2r - 3)}{r(1-m)}\pi_{r-1} - \frac{(\alpha + r - 1)(k - r + 2)m}{r(r-1)(1-m)}\pi_{r-2}\,, \tag{17}$$

$$\pi_0 = \sum_{s=0}^{\infty} \alpha_{(s)}(-k)_{(s)}m^s/s!(\alpha+\beta)_{(s)} \quad\text{and}\quad \pi_1 = -\frac{d\pi_0}{dm}.$$

For the particular cases of the above distribution, we may refer to Gurland [1958]. The recurrence relation for factorial cumulants is

$$K_{[r]} = \lambda\alpha_{(r)}k^{(r)}m^r/(\alpha+\beta)_{(r)}\,. \tag{18}$$

(ii) Poisson-Power series distribution:

$$w(z) = \sum_{i=0}^{\infty} a_i z^i, \tag{19}$$

where a_i's are constants such that $w(z)$ is convergent for some z. The recurrence relation for probabilities is

$$P_r = \lambda \sum_{s=1}^{r} s a_s P_{r-s}/r \quad\text{with}\quad P_0 = \exp\{\lambda(a_0 - 1)\}. \tag{20}$$

The above distribution was first given by Maritz [1952].

Type B: Let the distribution of egg masses be a general binomial with p.g.f. $(1 - p + pz)^n$. Then the distribution of the larvae over the whole field is $[1 - p + pw(z)]^n$, i.e. the Type B distribution with $h(z) = \{1 - p + pw(z)\}$. When $n = 1$, this leads to the distribution with p.g.f. $w(z)$ with an addition (or subtraction) of zeros. If $n > 1$ or $n < 0$, this can be regarded as n-th confluent of $w(z)$.

Some important distributions with their recurrence relations are:

(i) G. Binomial-Hypergeometric distribution: Here $w(z)$ is the same as given in (15). This was first stated by Gurland [1958] in the special case when $n < 0$, $p < 0$. The recurrence relation in probabilities is

$$P_r = p \sum_{s=1}^{r} (ns - r + s)\pi_s P_{r-s}/ra_0\,, \tag{21}$$

where $P_0 = a_0^n$, $a_0 = 1 - p + p\pi_0$ and π_i's are defined in (17).

(ii) G. Binomial-G. Binomial distribution: Here $w(z) = (1-m+mz)^k$. Hence, $\pi_i = m(k - i + 1)\pi_{i-1}/i(1 - m)$, with $\pi_0 = (1 - m)^k$, $a_0 = 1 - p + p\pi_0$ and $P_0 = a_0^n$. The recurrence formula has the same form as in (21).

(iii) G. Binomial-Poisson distribution: Here $w(z) = \exp\{\lambda(z - 1)\}$. Hence, $\pi_i = \lambda^i\pi_0$, $\pi_0 = \exp(-\lambda)$, $a_0 = 1 - p + p\pi_0$ and $P_0 = a_0^n$. The recurrence formula has the same form as in (21). When $n = 1$ and $p\{1 - \exp(-\lambda)\} = \theta$, this was called by A. C. Cohen [1960] an extension of a truncated Poisson distribution.

Type C: Let the distribution of egg masses be a logarithmic distribution with p.g.f. $\log(1 - \lambda z)/\log(1 - \lambda)$. Then the distribution of a larva over the whole field will be $\log\{1 - \lambda w(z)\}/\log(1 - \lambda)$, i.e. the Type C distribution with $h(z) = 1 - \lambda w(z)$. The important distributions are obtained by considering $w(z)$ as hypergeometric function G. Binomial, Power-series etc.

Choice of a distribution under the condition of no migration:

Let us suppose that the different sites of a colony are distinct and countable, and let there be no migration between sites of a colony. Assuming the same probabilities of arriving at a particular site by any organism, the distribution of r (when r is fixed) organisms in a particular site is Binomial. Now let us suppose that one or more organisms arriving at the colony follow the truncated Negative Binomial law. (This may be true under the wide applicability of the Negative Binomial in biological data, e.g., Bliss [1953], Evans [1953]). Then, it is easy to show that the p.g.f. of the organisms in a particular site (when there is no migration) is

$$1 - \theta + \theta\{(1 + m_1 - m_1 z)^{-k_1} - (1 + m_1)^{-k_1}\}/\{1 - (1 + m_1)^{-k_1}\}$$

where $0 < \theta < 1$, $m_1 > 0$, $k_1 > 0$; i.e. $1 - p + p(1 + m_1 - m_1 z)^{-k_1}$ for $0 < p < \{1 - (1 + m_1)^{-k_1}\}^{-1}$, $m_1 > 0$ and $k_1 > 0$. Now suppose that the independent results of n sites are combined together. Then the p.g.f. of the distribution of organisms is

$$[1 - p + p(1 + m_1 - m_1 z)^{-k_1}]^n$$

which is a particular case of G. Binomial-G. Binomial. The above distribution can be named as Binomial-Negative Binomial.

5. METHODS OF ESTIMATION FOR G. BINOMIAL-G. BINOMIAL

Here the p.g.f. is $[1 - p + p(1 - m + mz)^k]^n$.

Method of moments: Let $T = K_{[4]}/K_{[2]}^2$, $R = K_{[3]}K_{[1]}/K_{[2]}^2$ and

$S = K_{[1]}^2/K_{[2]}$. The approximate value of n is obtained from

$$n^2(TS - 2R^2 + R) + nS(TS - 6R + 6) + S^2(R - 2) = 0.$$

Then k, m and p are estimated from

$$k = \{S^2 + nS + n^2(2 - R)\}/n\{n(1 - R) - S\},$$

$$m = (nK_{[2]} + K_{[1]}^2)/(k - 1)nK_{[1]} \quad \text{and} \quad p = K_{[1]}/knm. \tag{22}$$

Method of maximum likelihood when n and k are fixed:

Similar to Sprott's results [1958], we have the maximum-likelihood equations as

$$\bar{r} = nk\hat{p}\hat{m} \quad \text{and} \quad L(\hat{m}) = \Sigma a_r G(r) - N = 0 \tag{23}$$

with $L'(\hat{m}) = (d/d\hat{m})L(\hat{m}) = \sum a_r G(r)[\hat{m}^{-1}\{1 - k(1 - \hat{p})^{-1}\} - \{1 + (1 - \hat{m})[\hat{m}k(1 - \hat{p})]^{-1}\hat{m}^{-1}\bar{r}\Delta\, G(r)\}]$, where $\bar{r}$ is the sample mean, a_r is the frequency at the r-th count, N is the total frequency, $G(r) = (r + 1)\hat{P}_{r+1}/\hat{P}_r\bar{r}$, $\Delta\, G(r) = G(r + 1) - G(r)$ and $\hat{m}$, $\hat{p}$ are maximum likelihood estimates.

From the first approximate estimates $\hat{p}'$, $\hat{m}'$, the new corrected values $\hat{p}''$, $\hat{m}''$ are estimated from

$$\hat{m}'' = \hat{m}' - \{L(\hat{m}')/L'(\hat{m}')\} \quad \text{and} \quad \hat{p}'' = \bar{r}/nk\hat{m}''. \tag{24}$$

Sample zero frequency when n and k are fixed:

Here the estimates p and m are obtained from

$$a_0 = N[1 - p + p(1 - m)^k]^n \quad \text{and} \quad \bar{r} = nkpm. \tag{25}$$

It may be noted that when $n = 1$, the two equations in (25) are the same as those in (23).

6. EXAMPLE

In order to illustrate how the above discussion can help an experimenter in the field of curve fitting, we fit here Binomial-Negative Binomial to the data in Distribution 1 of MacGuire *et al.* [1957, Appendix]. From the data, the first four factorial cumulants are $K_{[1]} = 2.5900156$, $K_{[2]} = 0.6877630$, $K_{[3]} = 0.0218497$, $K_{[4]} = 0.9080044$.

Hence, on taking $n = 1$, the solution of k, by the method of moments, correct to first decimal place is -12.0. Then the various estimates of m and p are $m = -0.2204963$, $p = 0.9788582$ by the method of zero-cell frequency or maximum-likelihood and $m = -0.2196584$, $p = 0.982591$ by the method of moments.

The fits are shown in Table 1. The fits of other distributions are given along side for reference purposes only.

TABLE 1.

Count per plot	Obs. fre-quency	Binomial-Neg. Binomial		Negative Binomial [1957]	Poisson; Binomial [1957]	Poisson Power Series[3]
		by M.L. or zero fr.	by method of moments			
0	355	355.000	346.445	324.30	341.84	339.072
1	600	622.478	628.153	660.37	644.37	645.036
2	781	730.994	735.340	734.06	728.03	730.150
3	567	616.306	618.023	610.45	609.14	610.886
4	441	417.545	417.393	408.82	415.60	416.079
5	245	241.296	240.550	236.54	242.72	242.339
6	135	123.567	122.747	122.49	125.17	124.532
7	42	57.406	56.846	58.12	58.20	57.655
8	17	24.632	24.315	25.68	24.76	24.417
9	11	9.395	9.731	10.70	9.75	9.567
10[2]	11	7.301	5.457	4.47	5.42	5.267
χ^2 with ν d.f.		19.184 7.d.f.	20.294 7.d.f.	34.52 8.d.f.	25.52 8.d.f	25.939 8.d.f.

[2] Expected frequencies are from 10 and above

[3] Probabilities are calculated from the p.g.f. $\exp\{-(a+b)+az+bz^2\}$ where $a = 1.9022526$ and $b = 0.3438815$.

7. ASYMPTOTIC EFFICIENCIES

Here we give the asymptotic efficiencies for the various methods of estimation for m and p only in a G. Binomial-G. Binomial distribution.

The determinant of the information matrix up to order N^{-1} for the maximum likelihood estimates m and p is

$$D_{m,p} = N^2n^2[npk\{k(1-p)+(1-m)m^{-1}\}R - \{k(1-p)-1\}^2]/(1-p)^2, \tag{26}$$

where $R = -1 + \sum G^2(r)P_r$ and $G(r) = (r+1)P_{r+1}/nkpmP_r$.

The determinant of the covariance matrix up to order N^{-1} for the moment estimates of m and p is

$$D = (K_2K_4 + 2K_2^3 - K_3^2)/[Nk(k-1)nmK_1]^2, \tag{27}$$

where

$$K_1 = k\,n\,m\,p, \qquad K_2/K_1 = 1 + (k-1)m - (K_1/n),$$

$$K_3/K_1 = 1 + 3(k - 1)m + (k - 1)(k - 2)m^2 - 3(K_2/n) - (K_1^2/n^2)$$

and

$$K_4/K_1 = 1 + 7(k - 1)m + 6(k - 1)(k - 2)m^2 + (k - 1)(k - 2)(k - 3)m^3 - 4(K_3/n) - 6(K_1K_2/n^2) - (K_1^3/n^3) - 3(K_2^2/nK_1).$$

The asymptotic efficiency in the restricted sense is

$$E_1 = 1/D\ D_{m,p}\ . \qquad (28)$$

The determinant of the covariance matrix up to order N^{-1} for the sample zero frequency estimates of m and p is

$$D(m, p) = m^2a_0^2\{a_1(1 - P_0) - P_0K_1\}/N^2a_2^2n^2K_1P_0\ , \qquad (29)$$

where $a_0 = 1 - p + p(1 - m)^k$, $a_1 = K_2/K_1$, $P_0 = a_0^n$ and $a_2 = 1 - (1 - m)^k - km(1 - m)^{k-1}$. The asymptotic efficiency with respect to the method of moments is

$$E_2 = D/D(m, p). \qquad (30)$$

It may be noted that $E_2 = E_1^{-1}$ when $n = 1$.

TABLE 2

ASYMPTOTIC EFFICIENCY E_2 RELATIVE TO THE METHOD OF MOMENTS WHEN $n = 1$ AND $k = -1$.

		m	
$p\{1 - (1 - m)^k\} = \theta$	−0.5	−1	−2
0.3	1.492	1.857	2.193
0.6	1.611	2.125	2.630
0.9	2.444	4.000	4.944

8. ACKNOWLEDGEMENT

The authors thank S. K. Katti for his valuable help in preparation of this paper.

REFERENCES.

Beall, G. and Rescia, R. R. [1953]. A generalisation of Neyman's contagious distributions. *Biometrics 9*, 354–86.

Bliss, C. I. [1953]. Fitting a negative binomial distribution to biological data. *Biometrics 9*, 176–200.

Cohen, A. C. [1960]. An extension of a truncated Poisson distribution. *Biometrics 16*, 446–50.

Evans, D. A. [1953]. Experimental evidence concerning contagious distributions in ecology. *Biometrika 40*, 186–221.

Feller, W. [1943]. On a general class of contagious distributions. *Ann. Math. Stat. 14*, 389–400.

Gurland, J. [1958]. A generalized class of contagious distributions. *Biometrics 14*, 229–49.

Gurland, J. [1957]. Some interrelations among compound and generalized distributions. *Biometrika 44*, 265–68.

Maritz, J. S. [1952]. Note on a certain family of discrete distributions. *Biometrika 39*, 196–98.

MacGuire, J. U., Brindley, T. A. and Bancroft, T. A. [1957]. The distribution of European corn borer larvae *Pyrausta Nubilalis* (Hbn.) in field corn. *Biometrics 13*, 65–78.

Neyman, J. [1939]. On a new class of contagious distribution applicable in entomology and bacteriology. *Ann. Math. Stat. 10*, 35–57.

Skellam, J. G. [1952]. Studies in statistical ecology, I. Spatial pattern. *Biometrika 39*, 346–62.

Sprott, D. A. [1958]. The method of maximum likelihood applied to the Poisson-Binomial distribution. *Biometrics 14*, 97–106.

FURTHER CONSIDERATION OF METHODOLOGY IN STUDIES OF PAIN RELIEF

PAUL MEIER
Department of Statistics, University of Chicago, Chicago, Illinois, U.S.A.

AND

SPENCER M. FREE, JR.
Smith, Kline, and French Laboratories. Philadelphia, Pennsylvania, U.S.A.

INTRODUCTION

In an earlier paper [1] we challenged the prevalent view that in comparisons of pain relieving drugs it is always desirable to have "each patient act as his own control," i.e., to test more than one drug on each patient and to estimate treatment contrasts from within-patient differences. We sought data of other investigators to compare with our own, but we were unable to find references which gave sufficient detail to permit investigation of the merits of alternative designs and analyses. It seemed desirable, therefore, to put our own data on record so as to facilitate discussion of the issues raised. This we did, and the response has been gratifying.

This paper presents and discusses some of the suggestions and comments of those who wrote to us. We are particularly indebted to Frederick Mosteller, Department of Statistics, Harvard University, and to Robert Curnow, A. R. C. Unit of Statistics, University of Aberdeen, Aberdeen, Scotland, who, in addition to making numerous thoughtful comments, performed their own analyses of our data.

A SIMPLE EXTENSION OF THE MODEL

As stated in our earlier paper, this study of relief of post-operative pain was designed as an incomplete block experiment. Each patient was given a dose of a test drug, and the dose was repeated when the patient again complained of severe pain. The number of hours of "greater than 50 percent pain relief" cumulated over both intervals was the measure of drug efficacy. A second drug was tested on each patient in the same way, starting at the time when the patient again complained of severe pain. Three drugs, two levels of a new drug and one level of Demerol, were under study.

It was found that the total hours of relief achieved when a given drug was the second administered was larger on the average than when that drug was administered first. This is consistent with the presumption that the pain decreases with time after operation. However, it invalidates a straightforward least squares analysis based on the model

$$\text{hours of relief} = \text{grand mean} + \text{patient effect} \\ + \text{drug effect} + \text{random error.}$$

We considered the possibility of extending the analysis to include a simple time-period effect, representing the average differences in relief between the second and first periods, but for two reasons we did not pursue it. Firstly, and most importantly, it seemed to us that such a model would be unlikely to be an accurate reflection of the true situation. For example, if the duration of drug effect increases with time after operation, the period effect should be greater when the second drug is given after a potent drug than when it is given after a weak drug. Secondly, our object in the first paper was to compare *simple* alternatives for the design and analysis of studies to evaluate analgesics, and a procedure requiring adjustment for time period as well as patient effects did not seem attractive as a procedure to recommend to clinicians for routine screening of new drugs. (In this connection one should remember that it is almost always necessary to discontinue study of some patients, often for reasons unrelated to drug response. Thus, balance in design is hardly ever achieved.)

Curnow extended the least squares analysis of our data to take account of such a time-period effect. The model becomes

$$\text{hours of relief} = \text{grand mean} + \text{patient effect} + \text{drug effect} \\ + \textit{time period effect} + \text{random error}$$

and the corresponding analysis is given in Table I. He stated that the differences due to the drug which follows a potent one being administered rather later than one which follows a weak one were small and had little effect on the estimates of drug differences. Thus the internal evidence of the experiment itself does not seem to confirm the fears expressed above, and the least squares analysis that includes a time-period effect appears to yield a reasonable description of the data. In particular, in both the inter- and intra-block analyses, the between-group residual mean square is now not larger than the corresponding within-groups mean square.

Curnow went on to point out that the inter-block error is not much larger than the intra-block error, and that one might on this ground justify an analysis *ignoring patient differences*, as shown in Table II.

TABLE I

LEAST SQUARES INTRA-BLOCK ANALYSIS

Source	D.F.	S.S.	M.S.
Blocks	42	607.256	
Periods and Drugs	3	378.015	
Periods ignoring drugs	1	261.628	
Drugs adjusted for periods	2	116.387	58.194
Between Groups Residual	3	3.051	1.017
Within Groups Residual	37	428.945	11.593
Total	85	1417.256	

DRUG COMPARISONS

Comparison	Mean Difference	Variance	t
D vs. T_1	3.421	1.178	3.15
D vs. T_3	2.037	1.061	1.98
T_3 vs. T_1	1.384	1.025	1.35

TABLE II

LEAST SQUARES ANALYSIS IGNORING PATIENT DIFFERENCES

Source	D.F.	S.S.	M.S.
Periods and Drugs	3	434.368	
Periods ignoring drugs	1	261.588	
Drugs adjusted for periods	2	172.780	86.390
Remainder	82	982.888	11.986
Total	85	1417.256	

DRUG COMPARISONS

Comparison	Mean Difference	Variance	t
D vs. T_1	3.548	0.874	3.80
D vs. T_3	1.861	0.831	2.04
T_3 vs. T_1	1.687	0.816	1.87

This observation suggests that in future experiments conducted under similar conditions we should *not* arrange matters to eliminate patient effects, if it will cost us much to do so.

Our own calculations and scatter diagrams confirm Curnow's results. However, in view of our first argument against the use of a simple time-period effect we found them surprising and made some further investigation. We take note first that the fact that a model "fits", in the sense that certain internal checks are satisfied, does not insure its correctness. If, for example, second-dose relief were more variable than first-dose relief, an examination of residuals would not in this case establish that fact; the estimated first-dose residual for a given patient has precisely the same magnitude as the estimated second-dose residual for that patient when we use the present model. The second-dose relief scores are, in fact, more variable than are those for first doses. This may be observed by examination of Table 1b in [1] and it is reflected in the fact that the *within patient* residual variance (Table 3a), which excludes the component of variability due to patient differences, is actually larger than the *between patient* variance for first dose relief (Table 2).

If the facts are in accord with the assumption that the increment in relief with increasing time after administration is proportional to time after operation, the effect of the model having only a simple time-period effect would be to assign a part of the drug difference to the time-period effect and thus to reduce the measured differences between drugs, as compared to an analysis for the first period only. In fact, the difference between the most and least potent drugs (D and T_1) estimated by use of the extended model is less by about 15 percent than the estimate obtained from analysis of the first period only.

Granted, then, that the model with a simple time-period effect fits the data insofar as intra-block checks are concerned, what may we expect to gain or lose if we go ahead and use it? We do have some evidence which tends to contradict the assumptions of this model, but the effects do not seem large. We have some slight evidence that we may lose discriminating power by misinterpreting drug differences as time-period effects, but, as judged from these data, the reduction in variance may well compensate. Insofar as testing (i.e., deciding which drug is better) is concerned, or estimating potency in an experiment with a full range of standards, the analysis based on this model should give essentially valid results.

In any event, apart from the question of the best analysis for these data, the smallness of the between patients component in the extended analysis suggests that we would have achieved at least equal precision for drug comparisons if we had kept each patient on a single drug throughout the experimental period. Had we done so, we would have had, in addition to possible benefits in precision, the comfort of closely imitating the ordinary clinical situation, and the advantage of an unbiased estimate of a clearly interpretable measure of effectiveness along with an unbiased estimate of its variance. We might also have had the cooperation of several more surgeons, some of whom, understandably, object to studies which require administration of several coded drugs to each patient.

A REGRESSION MODEL

In furtherance of the point of view that the cost of statistical work is generally small compared to the cost of gathering clinical data, Mosteller suggested that the data be analyzed in accordance with a model which might give a fairly realistic appraisal of the time-trend effect. In particular, he suggested that we not cumulate the two doses of each drug but analyze the data for single doses according to the following model.

$$y_{imd} = \alpha_i + \beta_m + \gamma_m t_{imd} + \epsilon_{imd} ,$$

where

α_i = effect of i-th subject,
β_m = effect of medication m when given at time zero,
γ_m = regression of effect of medication m on time of administration,
t_{imd} = time at which individual i receives the d-th dose of medication m,
ϵ_{imd} = random error.

This analysis was carried out, using the times of drug administration (not shown, but available upon request) in addition to the data already presented. It was found that the regression coefficients γ_m were quite close to one another and, if we restrict the model to the case of equal γ_m , the least squares estimate of this common value is 0.20, corresponding to an additional hour of drug relief for every five hour interval between operation and drug administration. Using this restricted model and doubling the estimated effects to make these results comparable to those for our other analyses, we have the estimates of treatment differences shown in Table III.

Assuming for now the correctness of the regression model, we see that these estimates are quite close to those obtained in the analysis using only the first drug for each patient—Table 2 of [1]. However, the variances are reduced by about one-third, so that the regression

TABLE III

Drug Comparisons from Regression Model Analysis

Comparison	Mean Difference	Variance	t
D vs. T_1	4.066	0.824	4.48
D vs. T_3	2.530	0.732	2.96
T_3 vs. T_1	1.534	0.708	1.82

analysis, using all the data, appears to have about 50 percent greater precision than the "first drug only" analysis, which uses only half the data.

Comparing the apparent precisions provided by the above analyses, we see that Curnow's intra-block analysis, adjusting for patients and time periods (Table I) appears comparable in precision to the "first drug only" analysis. Curnow's analysis ignoring patient effects (Table II) appears almost equal in precision to the analysis based on the regression model.

OTHER COMMENTS AND ERRATA

A third correspondent, Irwin Bross, Roswell Park Memorial Institute, Roswell Park, New York, raised several points, in part overlapping those above. In addition he pointed out that the results of the "least squares" analysis (Table 4a in [1]) differ appreciably from the analysis based on simple linear combinations of intra-individual contrasts. Since the design is nearly balanced, closer agreement might be expected, even though our "least squares" analysis fails to allow for time-period effects. In fact, contrary to the suggestion in [1], the "linear combinations" analysis is not really comparable to the "least squares" analysis. The reason is that in estimating, say, $D - T_1$, no account was taken of the information about this difference given by contrasting the direct estimate of $D - T_3$ with the estimate of $T_1 - T_3$. If we take an average of the directly observed difference, $D - T_1$, with the contrast obtained by combining $D - T_3$ with $T_1 - T_3$, weighting inversely as the estimated variances, and proceed similarly for the other comparisions, we get Table IV, which agrees much more closely with the "least squares" analysis (Table 4a in [1]). As might be expected, the "extended least squares" analysis which does allow for time-period effects (Table I) gives results in excellent agreement with Table IV.

This revision leads in turn to a more precise combined intra- and inter-block analysis in place of that shown in Table 3c in [1]. The result (not given) is quite close to the "least squares" combined analysis.

TABLE IV

DRUG COMPARISONS FROM ANALYSIS OF WITHIN PATIENT CONTRASTS

Comparison	Mean Difference	Variance	t
D vs. T_1	3.449	1.206	3.18
D vs. T_3	2.051	1.073	1.98
T_3 vs. T_1	1.398	1.031	1.38

Both Curnow and Bross pointed out that erroneous entries are shown in Table 4c in [1]. This table should be replaced with Table 4c (revised) as shown.

TABLE 4c (revised)

COMBINED INTRA- AND INTER-BLOCK ANALYSIS FOR DRUG DIFFERENCES

$$w = \frac{1}{1.174} = 0.8518 \qquad w' = \frac{1}{3.720} = 0.2688$$

Calculation of Average Mean Difference: Drug D vs. Drug T_1

$$\frac{(w)(3.150) + (w')(3.900)}{w + w'} = 3.330.$$

Calculation of Variance of Difference

$$\frac{1}{w + w'} = \frac{1}{0.8518 + 0.2688} = 0.892.$$

DRUG COMPARISONS

Comparison	Mean Difference	Variance	t
D vs. T_1	3.330	0.892	3.52
D vs. T_3	1.843	0.848	2.00
T_3 vs. T_1	1.442	0.828	1.58

DISCUSSION

The point which we wished to emphasize in [1] was that on account of the increase in duration of relief with time of administration, the use of simple intra-patient contrasts—"each patient his own control"—may not be optimal, and that an experiment in which each patient received only a single drug might be preferable.

The analyses proposed by Curnow and Mosteller have demonstrated that for this data a model taking account explicitly of the effect of time of drug administration will yield more efficient estimates than does the one-way analysis of the first period only. With respect to the analysis of the present data, both the Curnow and Mosteller models appear to fit quite well, and the estimates of the drug effects produced by their analyses are more precise than those derived from the analysis for the first period only.

It is not clear, however, whether the above finding should be construed as evidence supporting the need for more complex models and analyses than are currently in vogue, or whether instead it is evidence supporting our original viewpoint. If, as the evidence of Curnow's analysis suggests, patient differences are negligible, we would have no reason at all to use each patient as his own control. Were each patient given only one drug, the time-period effect of Curnow's model would be eliminated from treatment contrasts and the one-way analysis—using now the whole of the data—should be fully efficient.

Even if patient differences are not negligible, the advantages of simplicity, both of design and analysis, may outweigh a small gain in efficiency which could be achieved with the use of a more complex design and analysis. With our data, for example, the estimates of greatest apparent precision were those provided by the regression model, and the gain in precision compared to the analysis of the first period only was 50 percent. Such a gain is, of course, quite worth having, but it is not overwhelming, and it is easy to believe that had our design assigned one drug to each patient throughout the period of study, the one-way analysis might have equalled the present regression analysis in precision.

In conclusion, we must again emphasize that we do not claim that our findings apply to all kinds of pain studies. There may be many situations for which within-patient contrasts are far superior to between-patient contrasts. What we do claim is that no principal such as "each patient his own control" is entitled to the status of dogma. In some situations, at least, the simpler methods are better.

REFERENCES

[1] Meier, Paul, Free, S. M., and Jackson, G. L. [1958]. Reconsideration of methodology in studies of pain relief. *Biometrics 14*, 330–42.

FITTING A GEOMETRIC PROGRESSION TO FREQUENCIES

E. J. Williams
Division of Mathematical Statistics,
C. S. I. R. O., Canberra, Australia

SUMMARY

This paper discusses the interpretation of frequency data when the series of observed frequencies is assumed to arise from a population in which the expected frequencies form a geometric progression. Such situations occur in the study of steadily increasing insect populations and similar phenomena, where the common ratio of frequencies is related to the rate of growth of the population.

The estimation of the common ratio, and tests for the significance of departure from geometric trend, are discussed. Asymptotic formulae for the tail probabilities in large samples are determined.

The methods and tests of significance are illustrated by application to some experimental data.

I. INTRODUCTION

Observations are often recorded by classifying them into several classes and counting the frequencies of occurrence in each class. In general, the expected frequencies will be partly specified by theoretical considerations, but will often also depend on one or more unknown parameters.

The interpretation of such data then involves, firstly, testing its concordance with the assumed form of expected frequencies, and secondly, estimating the unknown parameters. Thus, for instance, in a steadily increasing population of organisms, the number of individuals expected in successive age-groups will be in geometric progression, the common ratio between the expected numbers representing not only the growth of the population but also the effects of mortality, migration and other factors. In such a study, one of the objects would be to test the concordance with the assumed geometric progression, and the other would be to estimate the common ratio accurately.

In general problems of this kind, the parameters may be estimated, and the accuracy of the estimates assessed, by the method of maximum likelihood. To test the fit of the model to the data, the χ^2 test is generally appropriate. If hypothetical values of the parameters have been specified, as occurs in many practical instances, their concordance with the estimates derived from the data can also be tested by means of χ^2.

In the particular case, with which this paper deals, of expected frequencies in geometric progression, there exist sufficient statistics for the two parameters involved, representing the general level of the frequencies and their geometric trend. This facilitates estimation and significance testing, although, since the equations of estimation are in general non-linear, iterative methods are required in the arithmetical work of calculating estimates.

The need to fit frequencies in this way arose from a method, devised by Dr. R. D. Hughes, Division of Entomology, C.S.I.R.O., of using the age distribution in an aphid population to study its rate of growth in the field. The device of using the proportions of individuals in the immature instars in a field population of aphids to give, firstly, the age distribution, and secondly, the rate of increase of the population, is the subject of a forthcoming paper by Hughes. In the present paper, some preliminary observations of instar distribution under controlled experimental conditions are discussed.

Chapman and Robson [1960] have considered the age distribution in a stationary population subject to constant mortality. As this formulation leads to a geometric progression of expected frequencies, many of their results anticipate results given in this paper. However, the general objects and scope of the two papers are different.

Aphid populations under uniform and favourable conditions increase at a constant rate, so that their expected numbers at equal intervals of time form a geometric progression. However, the numbers even of an initially small population quickly become too large to be counted accurately. An estimate of the rate of growth of the population may then be made from the age distribution. The immature aphid passes through four instars before reaching maturity. In a stationary population, the number expected in each instar is proportional to its duration. For some species of aphids, for instance *A. craccivora,* the average durations of the first three instars are probably equal under constant conditions, so that equal numbers will be expected in each instar in a stationary population. When the population is increasing at a constant rate, the numbers in each instar will approximate to a geometric progression.

If the common duration of each of the first three instars is c, and the growth-rate of the population is ρ, then the ratio of the number in each instar to that in the preceding one is approximately

$$e^{-c\rho}.$$

In this result, mortality has been neglected, since it has been shown that under favourable conditions mortality in the immature stages is negligible.

By taking a random sample from the aphid population and determining the number in each instar we can first check the validity of the assumption of uniform growth by testing whether the numbers are consistent with a geometric progression. Having satisfied ourselves on this point, we can then estimate the common ratio of the numbers. If c, the duration of each instar, is known, the growth-rate can then be determined.

If the durations of the different instars are different, the above method will be modified, and the estimation of the growth-rate is a little more complicated; in principle, however, such data will still provide information from which the growth-rate can be determined.

II. THE BASIC DISTRIBUTION

We consider a sample of N individuals classified into k classes, the observed frequency in class i being n_i, with expected value

$$E(n_i) = \lambda\mu^i \qquad (i = 0, 1, \cdots, k - 1).$$

If N is assumed to have a Poisson distribution, so has each of the n_i; the joint probability density is then

$$P(n_0, n_1, \cdots, n_{k-1}) = \prod (e^{-\lambda\mu^i}\lambda^{n_i}\mu^{in_i}/n_i\,!) = \frac{e^{-\lambda\phi_0(\mu)}\lambda^N\mu^T}{n_0\,!\,n_1\,!\cdots n_{k-1}\,!}, \quad (1)$$

where

$$T = n_1 + 2n_2 + \cdots + (k - 1)n_{k-1},$$

and

$$\phi_0(\mu) = 1 + \mu + \cdots + \mu^{k-1}.$$

From the form of the density it is apparent that N and T are a pair of sufficient statistics for the parameters λ and μ. Thus, questions of estimation may be referred to the joint distribution of N and T.

By summing the probability (1) over values of the n_i leading to the given values of N and T, we find the joint distribution of N and T as

$$P(N, T) = \frac{C(N, T)}{N!}\, e^{-\lambda\phi_0(\mu)}\lambda^N\mu^T, \quad (2)$$

where $C(N, T)$ is a numerical coefficient.

The number N has a Poisson distribution with mean $\lambda\phi_0(\mu)$;

$$P(N) = e^{-\lambda\phi_0(\mu)}[\lambda\phi_0(\mu)]^N/N!.$$

Hence for given N, the conditional density of T, which is independent of λ, is

$$P(T \mid N) = C(N, T)\mu^T/[\phi_0(\mu)]^N.$$

This is a special case of the distribution discussed by Noack [1950] and described by him as a power-series distribution. Our $[\phi(\mu)]^N$ takes the place of his $f(z)$. From this expression for the density we see that $C(N, T)$ is the coefficient of μ^T in $\phi_0(\mu)^N$. This fact enables the numerical coefficient to be determined directly.

This conditional density of T provides the basis for estimates and tests of significance about μ.

III. THE COMBINATORIAL FACTOR

The coefficient $C(N, T)$ is seen to be a generalization of the binomial coefficient (for the positive binomial, when $k = 2$; for the negative binomial, when $k = \infty$). It is the number of ways of allocating T like objects among N different cells, none of which may contain more than $k - 1$ objects (see Riordan [1958], page 104, where some recurrence relations and moment formulae are also given).

The coefficients are generated by the function

$$[\phi_0(\mu)]^N = \left(\frac{1 - \mu^k}{1 - \mu}\right)^N = \left[\sum \binom{N + i - 1}{N - 1}\mu^i\right]\left[\sum \binom{N}{j}(-\mu^k)^i\right].$$

On equating coefficients of μ^T we find

$$C(N, T) = \binom{N + T - 1}{N - 1} - \binom{N}{1}\binom{N + T - k - 1}{N - 1} + \binom{N}{2}\binom{N + T - 2k - 1}{N - 1} - \cdots \qquad (3)$$

The series has $1 + [T/k]$ terms. This expansion in terms of binomial coefficients is useful, especially if k is large, since the first few terms then give a close approximation.

Because of symmetry, $C(N, T) = C[N, (k - 1)N - T]$ so that results for $T > \frac{1}{2}(k - 1)N$ can be found from those for $T < \frac{1}{2}(k - 1)N$.

From the generating function or otherwise we may also deduce the recurrence relation

$$C(N, T) = C(N, T - 1) + C(N - 1, T) - C(N - 1, T - k)$$

which is useful for computation.

Hitherto we have been considering relations among the $C(N, T)$ for a fixed value of k. We now consider relations for different k, and shall indicate by a suffix the number of classes.

We may express the coefficients for k classes in terms of the coefficients corresponding to the factors of k, by means of the following reduction formula.

If $k = fg$,

$$\phi_{0k}(\mu) = \frac{1 - \mu^k}{1 - \mu} = \frac{1 - \mu^f}{1 - \mu} \cdot \frac{1 - \mu^{fg}}{1 - \mu^f} = \phi_{0f}(\mu) \cdot \phi_{0g}(\mu^f).$$

Hence

$$C_k(N, T) = C_f(N, T) + C_f(N, T - f)C_g(N, 1) + C_f(N, T - 2f)C_g(N, 2) + \cdots ,$$

a series of $1 + [T/f]$ terms. The terms of these series are all positive, and generally smaller than those of the series (3) of products of binomial coefficients. They may have some advantages for computation, provided the coefficients for f and g classes are known.

By differentiating the generating function with respect to its parameter we may prove recurrence formulae of the type

$$\begin{aligned} C(N, T) = {} & \frac{N}{(k - 1)N - 2T} \\ & \cdot [(k - 1)C(N - 1, T) + (k - 3)C(N - 1, T - 1) \\ & + \cdots - (k - 1)C(N - 1, T - k + 1)]. \end{aligned} \tag{4}$$

When $k = 3$, many particularly simple recurrence formulae may be established. Recurrence relations when $k = 3$ are

$$\begin{aligned} C(N, T) &= \frac{N}{T(2N - T)} \\ & \quad \cdot [(2N - 1)C(N - 1, T - 1) + 3(N - 1)C(N - 2, T - 2)] \\ &= \frac{1}{2T} [(2N - T + 1)C(N, T - 1) + 3NC(N - 1, T - 2)] \\ &= \frac{N}{N - T} [C(N - 1, T) - C(N - 1, T - 2)], \end{aligned}$$

the last being a particular case of (4). As k increases the recurrence formulae become more complicated.

IV. MOMENTS OF THE CONDITIONAL DISTRIBUTION OF T

The moments and cumulants of the conditional distribution of T are of general interest, and will also be of use when the determination of probabilities in the tails of the distribution is being considered.

Since the cumulants for a sample of N are simply N times those for a sample of 1, we shall consider a sample of 1, for which T equals X, the number of the class $(0, 1, \cdots, k-1)$ in which the observation falls.

Clearly, if

$$\phi_r(\mu) = \left(\mu \frac{d}{d\mu}\right)^r \phi_0(\mu) = \sum_{x=0}^{k-1} x^r \mu^x,$$

then the rth moment about zero is $\phi_r(\mu)/\phi_0(\mu)$. This result is equivalent to the results of Noack [1950], though expressed in a slightly different manner.

However, unless k is small, these expressions are not convenient for the computation of the central moments and cumulants, which are more simply found directly,

The moment-generating function of X is

$$E(e^{sX}) = \sum \mu^x e^{sx} / \sum \mu^x = \phi_0(\mu e^s)/\phi_0(\mu).$$

Thus the cumulant-generating function is

$$\begin{aligned} K(s) &= \log \phi_0(\mu e^s) - \log \phi_0(\mu) \\ &= -\log(1 - \mu e^s) + \log(1 - \mu^k e^{sk}) + \log(1 - \mu) - \log(1 - \mu^k) \\ &= \sum \frac{\mu^r}{r}(e^{rs} - 1) - \sum \frac{\mu^{rk}}{r}(e^{rsk} - 1). \end{aligned}$$

From this expansion we derive the particular results

$$\kappa_1 = \frac{\mu}{1-\mu} - \frac{k\mu^k}{1-\mu^k},$$

$$\kappa_2 = \frac{\mu}{(1-\mu)^2} - \frac{k^2\mu^k}{(1-\mu^k)^2},$$

$$\kappa_3 = \frac{\mu + \mu^2}{(1-\mu)^3} - \frac{k^3(\mu^k + \mu^{2k})}{(1-\mu^k)^3},$$

$$\kappa_4 = \frac{\mu + 4\mu^2 + \mu^3}{(1-\mu)^4} - \frac{k^4(\mu^k + 4\mu^{2k} + \mu^{3k})}{(1-\mu^k)^4}.$$

The form of the cumulant-generating function shows that, if u_k is a geometric variable with parameter μ^k, then $X + ku_k = u_1$. It also follows that, if $u_k(N)$ is a negative binomial variable with parameter μ^k and index N, then $T + ku_k(N) = u_1(N)$.

When $\mu = 1$, the distribution is the uniform discrete distribution, and the cumulants are expressible in terms of Bernoulli's numbers. We then have

$$K(s) = -\log\left(\frac{e^s - 1}{s}\right) + \log\left(\frac{e^{sk} - 1}{sk}\right)$$

$$= \frac{k-1}{2}s + (k^2 - 1)\frac{B_2}{2}\frac{s^2}{2!} + (k^4 - 1)\frac{B_4}{4}\frac{s^4}{4!} + \cdots ,$$

where $s/(e^s - 1) = 1 + \sum B_r s^r/r! = e^{Bs}$ in symbolic form. Hence $\kappa_r = (k^r - 1)B_r/r$; in particular, $\kappa_1 = (k - 1)/2$, $\kappa_2 = (k^2 - 1)/12$, $\kappa_3 = 0$, $\kappa_4 = -(k^4 - 1)/120$.

Relation between Cumulants Corresponding to μ and μ^{-1}.

When the series of frequencies is reversed, the ratio μ is replaced by its reciprocal. It therefore follows that, when μ is replaced by μ^{-1}, the odd cumulants other than κ_1 are simply changed in sign, and the even cumulants are unaffected. We have

$$\kappa_1(\mu^{-1}) = k - 1 - \kappa_1(\mu),$$

$$\kappa_r(\mu^{-1}) = (-1)^r\kappa_r(\mu) \qquad (r > 1).$$

These results may also be readily verified from the form of the cumulant-generating function.

Because of these facts, we need not consider values of μ outside the range (0, 1).

V. ESTIMATION OF THE RATIO OF FREQUENCIES

In the conditional distribution, T is a sufficient statistic for the parameter μ, the common ratio of the expected ferquencies. Thus the estimation of μ is straightforward. However, since the equation of estimation by the method of maximum likelihood is non-linear, iterative methods will be required to solve it.

The logarithm of the conditional probability of T, apart from terms independent of the parameter, is $L = T \log \mu - N \log \phi_0(\mu)$, and its first derivative with respect to μ is

$$(T/\mu) - (N\phi_1/\mu\phi_0). \tag{5}$$

Equating the derivative to zero gives the maximum likelihood estimator of μ. We indicate by an asterisk the maximum likelihood estimator and functions of it. Thus

$$T - (N\phi_1^*/\phi_0^*) = 0. \tag{6}$$

Since the equation is linear in T, it is clear that the estimator is to be found simply by equating T to its expectation: $T = N\kappa_1^*$, as may be verified from the results of the previous section.

The variance of the derivative (5) gives the information I_μ about μ in the sample. Being linear in T, the derivative has variance

$$V(T)/\mu^2 = N\kappa_2/\mu^2 = \frac{N}{\mu^2}\left(\frac{\mu}{(1-\mu)^2} - \frac{k^2\mu^k}{(1-\mu^k)^2}\right) = I_\mu \, .$$

This result is a special case of the results of Patil [1961], who investigates the estimation of the parameter of a generalized power-series distribution. In large samples, the reciprocal of I_μ approximates the variance of the estimate μ^*; that is, $V(\mu^*) = \mu^2/N\kappa_2$.

For purposes of calculating and tabulating the solutions of the maximum likelihood equation, we shall put

$$T/(k-1)N = v,$$

so that, for all k, $0 \leq v \leq 1$.

Then the estimating equation may be written

$$\kappa_1 = \frac{\mu}{1-\mu} - \frac{k\mu^k}{1-\mu^k} = (k-1)v. \tag{7}$$

An alternative form is

$$\frac{1}{1-\mu} - \frac{k}{1-\mu^k} = -(k-1)(1-v),$$

or

$$\frac{\mu^{-1}}{1-\mu^{-1}} - \frac{k\mu^{-k}}{1-\mu^{-k}} = (k-1)(1-v).$$

Thus, if μ^* is the root corresponding to v, then μ^{*-1} is the root corresponding to $1 - v$. In particular, if $v = \frac{1}{2}$, $\mu^* = 1$.

Since the roots corresponding to values of v exceeding $\frac{1}{2}$ are the reciprocals of roots corresponding to values of v less than $\frac{1}{2}$, we may henceforth confine attention to $v \leq \frac{1}{2}$, $\mu \leq 1$.

Equation (7) can be solved iteratively, once an approximate value of μ has been chosen. If v is not too near $\frac{1}{2}$, and k is not small, a first approximation is

$$\mu_1 = \frac{(k-1)v}{1+(k-1)v} = \frac{T}{T+N}\,, \tag{8}$$

and a second approximation is

$$\mu_2 = \mu_1 + k\mu_1^k(1-\mu_1)^2.$$

The difference between the two sides of (7) when an approximate value of μ is substituted represents $-\mu/N$ times (5), the first derivative of the likelihood. The adjustment to μ is given by the ratio of this difference to $-\mu I_\mu/N$. Then, approximately,

$$\mu^* = \mu\left[1 - \frac{\kappa_1 - (k-1)v}{\kappa_2}\right].$$

The substitution and adjustment may be repeated as often as required to give the desired accuracy.

If v is close to $\frac{1}{2}$, an approximation alternative to (8) is to be preferred. We put

$$\mu = e^{-\theta}, \qquad v = \tfrac{1}{2} - w.$$

Then equation (7) becomes

$$\frac{e^{-\theta}}{1 - e^{-\theta}} - \frac{ke^{-k\theta}}{1 - e^{-k\theta}} = (k-1)(\tfrac{1}{2} - w)$$

or, symbolically in terms of Bernoulli's numbers,

$$\frac{1}{\theta}(e^{B\theta} - e^{Bk\theta}) = (k-1)(\tfrac{1}{2} - w).$$

We then find

$$\frac{B_2\theta}{2!}(k^2 - 1) + \frac{B_4\theta^3}{4!}(k^4 - 1) + O(\theta^5) = (k-1)w,$$

whence

$$\theta = \frac{12w}{k+1} + \frac{144}{5}\frac{(k^2+1)w^3}{(k+1)^3} + O(w^5),$$

and

$$\mu^* = 1 - \frac{12w}{k+1} + \frac{72w^2}{(k+1)^2} - \frac{144}{5}\frac{(k^2+11)w^3}{(k+1)^3} + O(w^4). \qquad (9)$$

Solutions of the maximum likelihood equation for various values of k and v are given in Table 1. Once values of μ are given, it is easy to compute I_μ. As we shall see in Section X, the solution not only gives a point estimate of the ratio μ, and an approximate standard error based on the information function I_μ, but also gives a means of determining tail probabilities, needed in making significance tests and setting confidence limits.

VI. ESTIMATION OF λ

In general, for the problems considered in this paper, the actual value of λ, representing the size of the population (or rather, of the

TABLE 1

MAXIMUM LIKELIHOOD ESTIMATE OF μ [WITH ARGUMENT $v = T/(k-1)N$]

k	2	3	4	5
0.00	0	0	0	0
0.05	052632	092894	131334	167119
0.10	111111	178395	238358	291010
0.15	176471	261940	333333	392939
0.20	250000	346500	422530	483323
0.25	333333	434259	509668	567737
0.30	428571	527202	597388	649654
0.35	538462	627431	687922	731617
0.40	666667	737405	783468	815815
0.45	818182	860221	886484	904443
0.50	1.000000	1.000000	1.000000	1.000000

zero-class), is not of interest. However, for testing the significance of departure of the frequencies from a geometric progression, estimates of the expected frequencies in each class will sometimes be required. We shall then be interested in the simultaneous estimation of λ and μ, based on N and T. We therefore consider this question, rather than the estimation of λ alone.

The joint probability of N and T, as given in (2), is

$$\frac{C(N, T)}{N!} e^{-\lambda\phi_0(\mu)}\lambda^N\mu^T,$$

so that the logarithm of the likelihood, considered as a function of the two parameters, and with constant factors ignored, is $L = -\lambda\phi_0(\mu) + N \log \lambda + T \log \mu$. The derivatives are

$$\begin{aligned} L_\lambda &= -\phi_0(\mu) + (N/\lambda), \\ L_\mu &= [-\lambda\phi_1(\mu) + T]/\mu. \end{aligned} \tag{10}$$

The estimates are given when these derivatives are equated to zero, or when N and T are equated to their expected values. For direct solution, λ can be eliminated between the two equations, giving an equation for μ identical with (6). μ having been found, λ may then be found by substitution in (10).

For simultaneous solution, and also to give a measure of the information about λ and μ, we require the covariance matrix of the derivatives. Noting that N is a Poisson variate, with variance equal to its mean $\lambda\phi_0(\mu)$, that the covariance of N and T is $\lambda\phi_1(\mu)$ and that the variance of T is $\lambda\phi_2(\mu)$, we find for the covariance matrix of the derivatives

$$\mathbf{I} = \lambda\begin{bmatrix} \phi_0/\lambda^2 & \phi_1/\lambda\mu \\ \phi_1/\lambda\mu & \phi_2/\mu^2 \end{bmatrix}.$$

The inverse of $\mathbf{I}$, giving approximate variances and covariance of the estimates in large samples, is

$$\mathbf{I}^{-1} = \frac{1}{\lambda(\phi_0\phi_2 - \phi_1^2)}\begin{bmatrix} \lambda^2\phi_2 & -\lambda\mu\phi_1 \\ -\lambda\mu\phi_1 & \mu^2\phi_0 \end{bmatrix}.$$

From trial values λ, μ of the parameters we find improved estimates by means of the equations

$$\begin{bmatrix} \lambda^* \\ \mu^* \end{bmatrix} = \begin{bmatrix} \lambda \\ \mu \end{bmatrix} + \mathbf{I}^{-1}\begin{bmatrix} L_\lambda \\ L_\mu \end{bmatrix},$$

which give explicitly

$$\begin{aligned} \lambda^* &= \frac{N\phi_2 - T\phi_1}{\phi_0\phi_2 - \phi_1^2}, \\ \mu^* &= \mu\left[1 - \frac{N\phi_1 - T\phi_0}{\lambda(\phi_0\phi_2 - \phi_1^2)}\right]. \end{aligned} \tag{11}$$

Note that the trial value of λ does not appear in the equation for λ^*, since the estimating equations are linear in λ^*. The adjustments given by (11) may be repeated to give the accuracy required.

The expected frequency in class i is $\lambda\mu^i$; the variance of the estimated frequency may be determined, using $\mathbf{I}^{-1}$, as

$$\frac{\lambda\mu^{2i}}{\phi_0\phi_2 - \phi_1^2}(\phi_2 - 2i\phi_1 + i^2\phi_0).$$

The relative variance of the estimated frequency is therefore

$$\frac{\phi_2 - 2i\phi_1 + i^2\phi_0}{\lambda(\phi_0\phi_2 - \phi_1^2)}.$$

VII. TESTS OF DEPARTURE FROM PROPORTIONALITY

Before any use is made of data to which proportional frequencies have been fitted, it is necessary to test whether the observations depart significantly from the estimated frequencies.

Following Cochran [1954], we shall denote by X^2 the statistic used for testing the discrepancy between observed and expected frequencies, and by χ^2 the random variable with the familiar distribution, to which the distribution of X^2 approximates. We consider here the adequacy of the χ^2-approximation in testing X^2 for departure from expectation in the present problem.

Alternatively, we may use the fact that, because N and T are sufficient for λ and μ, the probability of a sample, conditional on N and T, is independent of the parameter values. An exact conditional test can thus be made, provided there are several possible samples corresponding to the given values of N and T. The probability of a sample of values $n_0, n_1, \cdots, n_{k-1}$ with $\sum n_i = N$ and $\sum in_i = T$ is

$$\frac{1}{\prod (n_i\,!)}\, e^{-\lambda\phi_0(\mu)}\lambda^N\mu^T.$$

Likewise the joint probability of N and T is

$$\frac{C(N, T)}{N!}\, e^{-\lambda\phi_0(\mu)}\lambda^N\mu^T,$$

where $C(N, T)$ is as defined in (2). Hence, the probability of the sample conditional on N and T is

$$\frac{N!}{C(N, T)\prod (n_i\,!)}. \tag{12}$$

By enumerating all possible samples and their probabilities we can decide the significance of an observed sample. Often the set of samples deemed to show significant departure from proportionality will be taken as a set of the samples with the smallest probabilities; however, in many problems, other criteria may be considered more appropriate. The use of X^2 has the advantage that it orders the samples according to a quantitative measure of departure from proportionality, whereas the probability attaching to a particular sample may be small merely because of the discontinuity of the distribution. A satisfactory arrangement would therefore be to order the samples by means of X^2, but to use the exact probabilities. In many cases the exact cumulated probabilities will not differ much from those given by the χ^2-distribution. The questions of calculation of probabilities and of choice of the significant set have been thoroughly discussed by Fisher [1950].

As an example of the exact calculation, we consider the following data in four classes:

	Observed	Expected
n_0	30	27
n_1	5	9
n_2	2	3
n_3	3	1

It will be seen that, because the numbers are restricted to four classes, the possibilities are much fewer in number than in other examples, such as that of the testing of deviations from a Poisson distribution, as discussed by Fisher. For this set, $N = 40$, $T = 18$, and it is readily verified that $\mu^* = \frac{1}{3}$, $\lambda^* = 27$. X^2 is easily calculated as

$$\frac{n_0^2}{27} + \frac{n_1^2}{9} + \frac{n_2^2}{3} + \frac{n_3^2}{1} - N = 6.444;$$

however, since the expected values in two of the classes are small, the tabulated distribution of χ^2 is not likely to be a good approximation to the distribution of X^2 for this sample. For this reason, as an example of method, the exact probabilities have been determined for the 37 possible samples having values $N = 40$, $T = 18$. The probabilities and the values of X^2 are given in Table 2.

We have $C(40, 18) = 220{,}495{,}831 \times 10^6$.

From the Table it is seen that the sample in question just attains significance at the 5 percent level; this is so whether the samples are ordered according to their probability or by X^2. The tabulated distribution would give a probability, for χ^2 with two degrees of freedom,

$$e^{-\frac{1}{2}X^2} = e^{-3.222} = 0.039866.$$

This shows that, for this example, the tabulated distribution underestimates the significance probability only slightly. On the other hand, for the sample

$$31 \quad 3 \quad 3 \quad 3$$

for which the probability accumulated according to X^2 is 0.008190,

$$X^2 = 8.593,$$

the probability of exceeding which is

$$0.013619.$$

The probability here is overestimated somewhat.

As a matter of interest, the mean and variance of for the distribution generated by this set are $E(X^2) = 1.9956$, $V(X^2) = 3.4223$, compared with the values of 2 and 4 respectively for the χ^2-distribution. The low variance will usually lead to overestimation of probabilities at the tails of the distribution and underestimation of significance, when referred to χ^2, though this effect will sometimes be masked by local irregularities of the distribution.

VIII. MOMENTS OF THE NON-PARAMETRIC DISTRIBUTION

We now show how the moments of the exact distribution of X^2

TABLE 2.

THE FREQUENCIES IN FOUR CLASSES FOR WHICH $N = 40$, $T = 18$

n_0	n_1	n_2	n_3	Probability	Cumulative probability	$27\ \chi^2$
34	0	0	6	.000000	.000000	1048
31	0	9	0	.000001	.000001	610
33	0	3	4	.000003	.000004	522
33	1	1	5	.000004	.000008	696
32	0	6	2	.000010	.000018	376
32	3	0	5	.000019	.000037	646
31	1	7	1	.000089	.000126	352
32	1	4	3	.000098	.000224	334
32	2	2	4	.000146	.000370	424
30	2	8	0	.000173	.000543	408
22	18	0	0	.000514	.001058	376
31	4	1	4	.000781	.001839	370
30	6	0	4	.000807	.002646	360
31	2	5	2	.000938	.003584	226
31	3	3	3	.002083	.005667	232
30	3	6	1	.003229	.008896	198
29	4	7	0	.003460	.012356	250
24	15	0	1	.004561	.016917	198
28	9	0	3	.005574	.022491	190
23	16	1	0	.006841	.029332	226
26	12	0	2	.009578	.038910	136
30	5	2	3	.009688	.048598	174
30	4	4	2	.012110	.060708	120
29	7	1	3	.013840	.074547	160
28	6	6	0	.023412	.097959	136
29	5	5	1	.029063	.127022	88
24	14	2	0	.034206	.161228	120
25	13	1	1	.038311	.199539	88
27	10	1	2	.046824	.246363	66
29	6	3	2	.048439	.294802	58
27	8	5	0	.070236	.365038	66
28	8	2	2	.075253	.440291	40
25	12	3	0	.083006	.523297	58
28	7	4	1	.100337	.623634	22
26	10	4	0	.105354	.728988	40
26	11	2	1	.114932	.843920	22
27	9	3	1	.156080	1.000000	0

may be determined. This distribution is independent of the population parameters, and depends only on the values of N and T.

We consider first the conditional moments of the n_i . From expression (12) we see that

$$\frac{C(N, T)}{N!} = \sum \frac{1}{\prod (n_i !)},$$

where the sum is taken over all sets of the n_i such that $\sum n_i = N$ and $\sum in_i = T$. Now the expected value of n_j is

$$\frac{N!}{C(N, T)} \sum \frac{n_j}{\prod (n_i !)}.$$

In the sum, the total of the arguments, with n_j replaced by $n_j - 1$, is $N - 1$, and the weighted sum of the arguments is $T - j$. It follows that $E(n_j) = NC(N - 1, T - j)/C(N, T)$. By a similar method it follows that, if $\sum a_i = A$, and $\sum ia_i = B$, then the general factorial moment is

$$E\left[\frac{\prod (n_i !)}{\prod (n_i - a_i)!}\right] = \frac{N!}{(N - A)!} \frac{C(N - A, T - B)}{C(N, T)}.$$

Thus in principle the joint factorial moments of the n_i can be determined provided we can regard the $C(N, T)$ as known. Some discussion of the asymptotic representation of $C(N, T)$ is given in Section X, but more needs to be known about these coefficients in general.

For the test of departure from expected frequencies we have

$$X^2 = \sum \frac{n_i^2}{\lambda^* \mu^{*i}} - N.$$

Now λ^* and μ^* are functions of N and T only, so are fixed for the conditional distribution. Hence X^2 may be simply regarded as a weighted sum of squares of the n_i. In particular, since

$$E(n_i^2) = \frac{N(N - 1)C(N - 2, T - 2i) + NC(N - 1, T - i)}{C(N, T)}$$

we have

$$E(X^2) = \frac{N}{\lambda^* C(N, T)}$$

$$\cdot \sum \frac{(N - 1)C(N - 2, T - 2i) + C(N - 1, T - i)}{\mu^{*i}} - N;$$

higher moments may be found similarly.

IX. TEST FOR TREND IN FREQUENCIES

Of particular importance is the test for the reality of any apparent trend in the frequencies—that is, whether the estimated value of μ differs significantly from unity. When N is large, the test may be made by comparing the difference $1 - \mu^*$ with its standard error, or, what is equivalent, testing $\frac{1}{2}(k - 1)N - T$ against its standard error.

When $\mu = 1$, the variance of μ^* reduces to $12/(k^2 - 1)N$, so that the test statistic, distributed approximately as χ^2 with 1 degree of freedom, is

$$(k^2 - 1)N(1 - \mu^*)^2/12. \tag{13}$$

More directly, the variance of T is

$$[(k^2 - 1)N]/12,$$

so that an alternative test statistic is

$$12(\tfrac{1}{2}(k - 1)N - T)^2/(k^2 - 1)N. \tag{14}$$

The statistics (13) and (14) are equivalent in large samples since, when μ^* is near to unity, we have

$$\mu^* = 1 - \frac{12(\frac{1}{2}(k - 1)N - T)}{(k^2 - 1)N} + O(N^{-1}),$$

as can be seen from (9).

The exact significance probability is given by the probability of a value of T less than or equal to that observed, which is easily computed directly when N is not large. This probability should be doubled, since both tails of the distribution are relevant to this test. With $\mu = 1$, this probability is simply

$$\sum_{r=0}^{T} C(N, r)/k^N = S(N, T)/k^N.$$

Now the generating function of $S(N, T)$ is

$$[\phi_0(\mu)]^N/(1 - \mu) = (1 - \mu^k)^N/(1 - \mu)^{N+1}.$$

From this representation we readily deduce that

$$S(N, T) = \binom{N + T}{N} - \binom{N}{1}\binom{N + T - k}{N} + \binom{N}{2}\binom{N + T - 2k}{N} - \cdots \tag{15}$$

which provides the most convenient means of computation of isolated values of the sum. On putting $T = (k - 1)N$ in (15) we have the interesting corollary

$$\binom{kN}{N} - \binom{N}{1}\binom{k(N - 1)}{N} + \binom{N}{2}\binom{k(N - 2)}{N} - \cdots = k^N.$$

Values of T less than $\frac{1}{2}(k - 1)N$ required for significance at the 5 and 1 percent levels of probability have been determined for various values of k and N, and are presented in Table 3. For values of N beyond the range of the Table, the large-sample χ^2-test may be used.

TABLE 3

EXACT TEST FOR EXISTENCE OF TREND

(I.E. OF NULL HYPOTHESIS $\mu = 1$)

5 PERCENT POINT OF T [OR $(k - 1)N - T$]

	k		
N	3	4	5
5	1	2	3
6	1	3	4
7	2	4	6
8	3	5	7
9	3	6	9
10	4	7	10
11	5	8	12
12	6	9	13
13	6	11	15
14	7	12	17
15	8	13	18

1 PERCENT POINT OF T [OR $(k - 1)N - T$]

	k		
N	3	4	5
5	0	0	1
6	0	1	2
7	1	2	4
8	1	3	5
9	2	4	6
10	3	5	8
11	3	6	9
12	4	7	11
13	5	8	12
14	5	9	14
15	6	11	15

X. EVALUATION OF TAIL PROBABILITIES

When N is large, the calculation of the probabilities of the tails of the distribution is troublesome. Blackwell and Hodges [1959] have given a method for expeditiously finding approximate tail probabilities. The method depends on a transformation of the distribution to a new distribution for which probabilities in the neighbourhood of the mean are to be determined. Daniels [1954] has given equivalent results, expressed explicitly for continuous probability densities, and Good [1957] has given similar results to the term of order N^{-2} of the leading term.

We briefly outline Blackwell and Hodges' results applied to the present distribution. We make use of the fact that the distribution of T is the N-fold convolution of the distribution of X as defined in Section IV.

The moment-generating function of $X - a$, where a is any constant, is $M(s) = E[e^{s(X-a)}]$.

Let s^* be the unique value of s that minimizes $M(s)$, and denote the minimum of $M(s)$ by $m(a)$; note that s^* is a function of a. Then we transform to a new variable Y, for which

$$P(Y = x) = \frac{P(X = x)e^{s^*(x-a)}}{m(a)}. \qquad (16)$$

It is readily seen that Y has in fact a probability density with the same range as X. Also, by differentiating the numerator of (16) with respect to s^* and summing, we find that $E(Y) = a$.

Before continuing with Blackwell and Hodges' method, we establish the interesting result that, for any distribution for which the sum of the values of X is a sufficient statistic for the parameter μ, the new variate Y has a distribution of the same form, but with parameter μ^*, the maximum likelihood estimate corresponding to the value $X = a$. This result is probably well known, though we have never seen it discussed. Thus we see that, for distributions of this commonly occurring class, the solution of the maximum likelihood equation is important not only for providing a point estimate, but also for evaluating the tail probabilities corresponding to the particular observed value.

If the sum of the values of X is sufficient for μ, the density of X may be written as

$$h(X)e^{Xg(\mu)}/f(\mu), \qquad (17)$$

where

$$f(\mu) = \sum_x h(x)e^{xg(\mu)}.$$

Then

$$M(s) = \sum_x h(x)e^{s(x-a)+xg(\mu)}/f(\mu),$$

so that, for any value of s whatever, a transformed density is given by

$$h(X)e^{X[s+g(\mu)]}/e^{as}M(s).$$

This density is clearly of the same form as that of X.

Now the maximum likelihood equation, corresponding to an observed value a, turns out to be

$$a - \sum_x xh(x)e^{xg(\mu)}/\sum_x h(x)e^{xg(\mu)} = 0,$$

where we have differentiated (17) with respect to $g(\mu)$ rather than μ itself. This is in fact $E(X \mid \mu = \mu^*) = a$. Now from (16) we derived that $E(Y) = a$. It follows therefore that the parameter of the distribution of Y is μ^*.

We can also express $m(a)$ in terms of the maximum likelihood estimator. On putting $x = a$ in (16), we see that $m(a) = P(X = a)/P(Y = a)$, or $P(a \mid \mu)/P(a \mid \mu^*)$.

For the particular distribution we have been considering, therefore

$$m(a) = \left(\frac{\mu}{\mu^*}\right)^a \left(\frac{\phi_0^*}{\phi_0}\right),$$

where we indicate by an asterisk a function of μ^*.

Blackwell and Hodges, by means of the transformation (16), and taking $Na = T$, show that

$$P(X_1 + X_2 + \cdots + X_N = T) = [m(a)]^N P(Y_1 + Y_2 + \cdots + Y_N = T),$$

and thence, that each side is equal to

$$\frac{[m(a)]^N}{\sqrt{2\pi N \kappa_2^*}} \left[1 + \frac{1}{8N}\left(\frac{\kappa_4^*}{\kappa_2^{*2}} - \frac{5}{3}\frac{\kappa_3^{*2}}{\kappa_2^{*2}}\right) + O(N^{-2})\right],$$

or

$$\left(\frac{\mu}{\mu^*}\right)^T \left(\frac{\phi_0^*}{\phi_0}\right)^N \frac{1}{\sqrt{2\pi N \kappa_2^*}} \left[1 + \frac{1}{8N}\left(\frac{\kappa_4^*}{\kappa_2^{*2}} - \frac{5}{3}\frac{\kappa_3^{*2}}{\kappa_2^{*3}}\right) + O(N^{-2})\right] \qquad (18)$$

where the κ_r^* are the cumulants of the distribution of Y; that is, functions of μ^*. The explicit terms in (18) give a satisfactory approximation to $P(T \mid N)$, provided N is large.

We note that the series in (18) depends only on μ^* (i.e., on T), and not on μ. Thus what we have obtained is an asymptotic expansion of $C(N, T)$. In fact,

$$C(N, T) = \frac{\phi_0^{*N}}{\mu^{*T}\sqrt{2\pi N \kappa_2^*}}\left[1 + \frac{1}{8N}\left(\frac{\kappa_4^*}{\kappa_2^{*2}} - \frac{5}{3}\frac{\kappa_3^{*2}}{\kappa_2^{*3}}\right) + O(N^{-2})\right].$$

Thus, $C(N, T)$ may be estimated by substituting the actual values of the cumulants, from the formulae given in Section IV.

As a check on the asymptotic formula, we estimate $C(N, T)$ for the values $k = 4$, $N = 15$, $T = 11$. We deduce $\mu^* = \frac{1}{2}$, $\phi_0 = 15/8$,

$$\kappa_2^* = 194/15^2 = 0.862222,$$

$$\kappa_3^* = 2842/15^3 = 0.842074,$$

$$\kappa_4^* = 1434/15^4 = 0.028326.$$

Hence

$$C(15, 11) \sim \sqrt{\frac{15}{2\pi\cdot 194}}\,\frac{(15/8)^{15}}{(\frac{1}{2})^{11}}\left(1 - \frac{1.805604}{120}\right) = 2{,}784{,}963.$$

The true value is

$$\binom{25}{15} - \binom{15}{1}\binom{21}{15} + \binom{15}{2}\binom{17}{15} = 2{,}784{,}600,$$

from which the estimate by the asymptotic formula is in error by 130 per million.

For significance testing, the cumulative tail probabilities are more important than the individual terms. These cumulative probabilities are also given by Blackwell and Hodges. They give

$$P(T \geq t) = \frac{[m(a)]^N}{\sqrt{2\pi N \kappa_2^*(1 - z)}}\left[1 + \frac{1}{8N}\left(\frac{\kappa_4^*}{\kappa_2^{*2}} - \frac{5}{3}\frac{\kappa_3^{*2}}{\kappa_2^{*3}}\right)\right.$$

$$\left. - \frac{z}{2N}\left(\frac{\kappa_3^*(1 - z) + \kappa_2^*(1 + z)}{\kappa_2^{*2}(1 - z)^2}\right) + O(N^{-2})\right]$$

$$= \frac{P(T = t)}{(1 - z)}\left[1 - \frac{z}{2N}\,\frac{\kappa_3^*(1 - z) + \kappa_2^*(1 + z)}{\kappa_2^{*2}(1 - z)^2} + O(N^{-2})\right]$$

where, for the distribution considered here, $z = \mu/\mu^*$. We note that, as is to be expected, the bracketed series in the cumulative probability depends on μ. The approximation is valid only for $\mu^* > \mu$, and is most effective when z is small—that is, when μ^* is large and the extreme tails of the distribution are being considered.

When we are determining the probability in the lower tail of the distribution, that is $P(T \leq t)$, so that $T < \frac{1}{2}(k - 1)N$, and $\mu^* < \mu$, we simply replace z by z^{-1} throughout the formula, and change the sign of κ_3^*.

As a test of the adequacy of this approximation, we apply it to the example for which the individual term was calculated above, namely $k = 4$, $N = 15$, $T = 11$.

Inserting the values found above, we have $z = 2\mu$,

$$P(T \leq 11) = \frac{P(T = 11)}{\left(1 - \frac{1}{2\mu}\right)} \cdot \left[1 - \frac{1}{60\mu} \left\{ \frac{\frac{-2842}{15^3}\left(1 - \frac{1}{2\mu}\right) + \frac{194}{15^2}\left(1 + \frac{1}{2\mu}\right)}{\frac{194^2}{15^4}\left(1 - \frac{1}{2\mu}\right)^2} \right\} + O(N^{-2}) \right].$$

We consider the adequacy of the approximation when $\mu = 1$, for which we can readily determine the exact probability. We have

$$P(T = 11) = 2{,}784{,}600/4^{15},$$

so the leading term in the approximate probability is

$$2P(T = 11) = 5{,}569{,}200/4^{15} = 0.005187.$$

For the first adjustment we find the factor $1 - 0.078223$ giving $5{,}133{,}560/4^{15} = 0.004781$. The exact value of the denominator is

$$S(15, 11) = \binom{26}{15} - \binom{15}{1}\binom{22}{15} + \binom{15}{2}\binom{18}{15} = 5{,}253{,}680,$$

giving the exact probability 0.004893.

We see that the first adjustment gives an approximation adequate for the purpose of assessing the significance probability. Since both tails of the distribution are relevant to any test of significance, the significance probability is 0.009786; $T = 11$ is thus the 1 percent point of the distribution.

By determining the tail probabilities corresponding to any given value of μ we can in principle set a confidence range for μ — the set of all μ which are not discordant with the observed values N and T.

XI. APPLICATION TO EXPERIMENTAL DATA

The data in Table 4 show numbers in each of the first three instars of immature cowpea aphids (*A. craccivora*), from samples drawn on six different occasions from the same population. Since the experimental conditions were uniform, and the average duration of each instar is the same (about 42 hours at 20°C), the expected numbers are in geometric progression. The Table also gives an overall estimate of the common ratio, 0.529820. In order to determine the growth-rate of the population,

TABLE 4

AGE-DISTRIBUTION OF IMMATURE COWPEA APHIDS (*A. craccivora*)

Sample	Instar			N	T	λ^*	μ^*	X^2
	I	II	III					
1	181	92	51	324	194	179.730	0.526015	0.11
2	148	78	42	268	162	147.737	0.531518	0.01
3	130	54	43	227	140	123.456	0.543412	4.07
4	70	37	21	128	79	69.580	0.543848	0.03
5	88	41	20	149	81	87.710	0.474049	0.13
6	85	42	28	155	98	82.857	0.558629	0.63
Total	702	344	205	1251	754	690.958	0.529820	2.14

we equate this ratio to $e^{-c\rho}$, c being the average duration of instar and ρ the growth-rate.

We then find

$$c\rho = -\log(0.529820) = 0.6352$$

so that

$$\rho = 0.6352/42 \text{ per hour} = 0.01512 \text{ per hour} = 0.363 \text{ per day.}$$

The population grows at the rate of $100(e^{0.363} - 1) = 44$ percent per day.

Table 4 also gives estimates of λ and μ for each sample, and X^2 with 1 degree of freedom measuring departure from the common ratio. Only the largest X^2-value, for sample 3, attains the 5 percent level of significance, so that there is no evidence for departure from hypothesis.

The variance of the estimate of μ from a sample of N is $\mu^2/V(T)$ which reduces when $k = 3$ to

$$\frac{\mu(1 + \mu + \mu^2)^2}{N(1 + 4\mu + \mu^2)}.$$

Using the overall estimate $\mu_0^* = 0.529820$ we find the variance of estimate to be

$$V(\mu^* \mid N) = 0.510813/N.$$

The consistency of the estimates of μ from the different samples may be tested by X^2.

For consistency,

$$X^2 = (0.526015^2 \times 324 + 0.531518^2 \times 268 + \cdots - 0.529820^2 \times 1251)/0.510813$$

$$= 1.84, \text{ with 5 degrees of freedom.}$$

The individual ratios differ no more than would be expected by chance.

After the manner of Section IX, formulae (13) and (14), an alternative statistic for testing consistency is based on comparison of the ratios T/N.

$$V(T/N) = \mu(1 + 4\mu + \mu^2)/N(1 + \mu + \mu^2)^2 = 0.549534/N$$

Hence

$$X^2 = (194^2/324 + 162^2/268 + \cdots - 754^2/1251)/0.549534$$
$$= 1.33.$$

REFERENCES

Blackwell, David and J. L. Hodges, Jr. [1959]. The probability in the extreme tail of a convolution. *Ann. Math. Stat. 30*, 1113–20.

Chapman, D. G., and D. S. Robson [1960]. The analysis of a catch curve. *Biometrics 16*, 354–68.

Cochran, W. G. [1954]. Some methods for strengthening the common χ^2 tests. *Biometrics 10*, 417–51.

Daniels, H. E. [1954]. Saddlepoint approximations in statistics. *Ann. Math. Stat. 25*, 631–50.

Fisher, R. A. [1950]. The significance of deviations from expectation in a Poisson series. *Biometrics 6*, 17–24.

Good, I. J. [1957]. Saddle-point methods for the multinomial distribution. *Ann. Math. Stat. 28*, 861–81.

Noack, Albert [1950]. A class of random variables with discrete distributions. *Ann. Math. Stat. 21*, 127–32.

Patil, G. P. [1961]. On homogeneity and combined estimation for generalized power series distribution and certain applications. *Biometrics 18,* to be published.

Riordan, John [1958]. *An Introduction to Combinatorial Analysis.* New York; Wiley.

COMPUTING PROCEDURES FOR ESTIMATING COMPONENTS OF VARIANCE IN THE TWO-WAY CLASSIFICATION, MIXED MODEL[1]

S. R. Searle,
New Zealand Dairy Board, Wellington, New Zealand,
AND
C. R. Henderson
Cornell University, Ithaca, N. Y., U. S. A.

INTRODUCTION

Methods for estimating variance components from unbalanced data are given in Henderson [1953] for both the random model and the mixed model. The calculations involved in the latter case are somewhat tedious and usually not computationally feasible for data having many classes. This paper outlines the analysis for unbalanced data from a two-way classification with one classification considered fixed. Simplified computing procedures are presented suitable for a large number of levels in the random classification and a reasonable number of levels of the fixed classification.

MODELS AND ESTIMATION

The model for the two-way classification can be taken as

$$x_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk} , \tag{1}$$

where x_{ijk} is the observation and μ is the general mean. α_i is the effect due to the i'th level of the α-class, β_j the effect due to the j'th level of the β-class, $\alpha\beta_{ij}$ the interaction and e_{ijk} a random error term. We will suppose that the number of α-classes in the data is a, the number of β-classes is b, that there are n_{ij} observations in the ij'th subclass, and that s sub-classes have observations in them. All terms except μ are taken as independent random variables in the random model with zero means and variances σ_α^2 , σ_β^2 , $\sigma_{\alpha\beta}^2$ and σ_e^2 respectively. The β-classification will be considered fixed in the mixed model, the variances involved being σ_α^2 , $\sigma_{\alpha\beta}^2$, and σ_e^2 , the interaction terms being random for the particular set of fixed effects occurring in the data. In this case we are assuming an underlying multivariate distribution with x_{ijk} having mean

[1] Parts of this paper are based on the thesis submitted by the first author in partial fulfillment of the requirements of the Ph. D. degree.

value $\mu + \beta$, and variance $\sigma_\alpha^2 + \sigma_{\alpha\beta}^2 + \sigma_e^2$: the covariance between x_{ijk} and $x_{ijk'}$ for $k \neq k'$ is $\sigma_\alpha^2 + \sigma_{\alpha\beta}^2$ and that between x_{ijk} and $x_{ii'k'}$ for $j \neq j$ and $k \neq k'$, is σ_α^2 .

Variance components can be estimated in both models by equating linear functions of sums of squares to their expected values, the sums of squares used being reductions in the total sum of squares due to fitting various elements of the model as if it were a fixed model. For example, fitting $(\mu + \alpha_i)$ results in matrix equations of the form

$$\mathbf{P}\hat{\boldsymbol{\alpha}} = \mathbf{y},$$

where $\mathbf{y}$ is the vector of α-class totals $x_{i..}$, and $\hat{\boldsymbol{\alpha}}$ is the vector of estimates of the α's. $\mathbf{P}$ is non-singular, a diagonal matrix of order a, of terms $n_{i.}$. Thus

$$\hat{\boldsymbol{\alpha}} = \mathbf{P}^{-1}\mathbf{y},$$

and the reduction in the total sum of squares due to fitting the α's is

$$R(\mu, \alpha) = \hat{\boldsymbol{\alpha}}\mathbf{y} = \mathbf{y}'\mathbf{P}^{-1}\mathbf{y}.$$

This is the usual uncorrected sum of squares, namely

$$R(\mu, \alpha) = \sum_i x_{i..}^2/n_{i.} \; . \tag{2}$$

Similarly

$$R(\mu, \beta) = \sum_j x_{.j.}^2/n_{.j} \; , \tag{3}$$

$$R(\mu, \alpha, \beta, \alpha\beta) = \sum_i \sum_j x_{ij.}^2/n_{ij} \; , \tag{4}$$

$$R(\mu) = x_{...}^2/n_{..} \; , \tag{5}$$

and

$$R(0) = \sum_i \sum_j \sum_k x_{ijk}^2 \; . \tag{6}$$

The variance components of the random model are estimated by equating the expected values to the observed values of differences among the above uncorrected sums of squares, namely (2)–(5), (3)–(5), (4)–(2)–(3) + (5) and (6)–(4). These cannot be used in the mixed model because, apart from (6)–(4), their expectations then contain functions of the fixed effects. The within-subclasses sum of squares, (6)–(4), can still be used to estimate the error variance as in the random model, and σ_α^2 and $\sigma_{\alpha\beta}^2$ can be estimated from the differences:

$$R(\mu, \alpha, \beta) - R(\mu, \beta),$$

and

$$R(\mu, \alpha, \beta, \alpha\beta) - R(\mu, \alpha, \beta),$$

whose expectations are free of terms in the fixed effects. $R(\mu, \alpha, \beta)$ is the reduction in the total sum of squares due to fitting α and β alone without the interaction terms. This reduction in sum of squares does not occur in the random model analysis and, when required for the mixed model analysis it usually involves a considerable amount of computing. Its relationship to the random model analysis will now be shown and a simplified computing procedure derived.

SIMPLIFIED EXPRESSION FOR R (μ, α, β)

Henderson *et al.* [1959] have shown that fixed effects in a mixed model can be estimated by treating the random effects as fixed and maximizing the joint distribution function of the observations and the random effects. The resulting equations for the two-way classification without interaction are

$$\begin{bmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{R} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \qquad (7)$$

where P is a diagonal matrix, of order a, with terms $n_{i.}$, $\mathbf{R}$ is a diagonal matrix of order $b - 1$ with terms $n_{.j}$, $j \neq b$, $\mathbf{Q}$ is a matrix of order a by $b - 1$ with terms n_{ij} , $j \neq b$, $\mathbf{y}$ is a vector of the $x_{i..}$ totals and $\mathbf{z}$ is a vector of the $x_{.j.}$ totals. $\mathbf{R}$, $\mathbf{Q}$ and $\mathbf{z}'$ have $b - 1$ columns because of omitting the equation for the last β, appropriate to imposing the constraint $\hat{\beta}_b = 0$ in order to have a unique solution. $R(\mu, \alpha, \beta)$ from these equations is

$$\begin{aligned} R(\mu, \alpha, \beta) &= (\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\beta}}') \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}, \\ &= (\mathbf{y}'\mathbf{z}') \begin{bmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{R} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}. \qquad (8) \end{aligned}$$

This expression for obtaining $R(\mu, \alpha, \beta)$ requires inverting a matrix of order $(a + b - 1)$, which is not feasible in many situations because a, the number of random classes in the data, is large. For example, the analysis of dairy production records in Searle and Henderson [1960] involved 688 herds (random) and 4 age groupings (fixed) so that a matrix of order 691 would need to be inverted for equation (8). However, because of its special form, it can be reduced to inverting a 3 × 3 matrix. Thus in the general case for the two-way classification the inversion of an $(a + b - 1)$ matrix can be reduced to inverting one of order $(b - 1)$ and this is a considerable reduction in the computing required especially when a, the number of random effects, is large, as is frequently the case.

The simplification of (8) proceeds as follows. Let

$$\begin{bmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{R} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{bmatrix}. \tag{9}$$

Then

$$\begin{aligned} \mathbf{A} &= \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{Q}\mathbf{D}\mathbf{Q}'\mathbf{P}^{-1}, \\ \mathbf{B} &= -\mathbf{P}^{-1}\mathbf{Q}\mathbf{D}, \\ \mathbf{D} &= (\mathbf{R} - \mathbf{Q}'\mathbf{P}^{-1}\mathbf{Q})^{-1}, \end{aligned} \tag{10}$$

$$\begin{aligned} R(\mu, \alpha, \beta) &= (\mathbf{y}'\mathbf{z}')\begin{bmatrix} \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{Q}\mathbf{D}\mathbf{Q}'\mathbf{P}^{-1} & -\mathbf{P}^{-1}\mathbf{Q}\mathbf{D} \\ -\mathbf{D}\mathbf{Q}'\mathbf{P}^{-1} & \mathbf{D} \end{bmatrix}\begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \\ &= \mathbf{y}'\mathbf{P}^{-1}\mathbf{y} + (\mathbf{z}' - \mathbf{y}'\mathbf{P}^{-1}\mathbf{Q})\mathbf{D}(\mathbf{z} - \mathbf{P}^{-1}\mathbf{Q}'\mathbf{y}). \end{aligned} \tag{11}$$

The matrix $\mathbf{P}$ is easily inverted, being diagonal, and the only non-diagonal matrix to invert is $\mathbf{D} = (\mathbf{R} - \mathbf{Q}'\mathbf{P}^{-1}\mathbf{Q})^{-1}$ of order $b - 1$.

The first term in the above expression is $R(\mu, \alpha)$, and the second can be derived from equations (7). Eliminating

$$\hat{\boldsymbol{\alpha}} = \mathbf{P}^{-1}(\mathbf{y} - \mathbf{Q}\hat{\boldsymbol{\beta}}),$$

gives

$$\hat{\boldsymbol{\beta}} = \mathbf{D}(\mathbf{z} - \mathbf{Q}'\mathbf{P}^{-1}\mathbf{y}). \tag{12}$$

This is easily computed and can then be used to obtain R_β, the reduction in the sum of squares due to fitting the β's in this no-interaction model, as

$$R_\beta = \hat{\boldsymbol{\beta}}'(\mathbf{z} - \mathbf{Q}'\mathbf{P}^{-1}\mathbf{y}). \tag{13}$$

Thus from (11) $R(\mu, \alpha, \beta)$ can be expressed as

$$R(\mu, \alpha, \beta) = R(\mu, \alpha) + R_\beta, \tag{14}$$

the first term of which is part of the random model analysis and the second term comes from (12) and (13).

EXPECTED VALUES

The expectation of $R(\mu, \alpha, \beta)$ can be found by using (8) and (9) and writing

$$\begin{aligned} R(\mu, \alpha, \beta) &= (\mathbf{y}'\mathbf{z}')\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{D} \end{bmatrix}\begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}, \\ &= (\mathbf{y}'\mathbf{A}\mathbf{y} + \mathbf{z}'\mathbf{B}\mathbf{y} + \mathbf{y}'\mathbf{B}\mathbf{z} + \mathbf{z}'\mathbf{D}\mathbf{z}), \\ &= \operatorname{tr}(\mathbf{A}\mathbf{y}\mathbf{y}' + 2\mathbf{B}\mathbf{z}\mathbf{y}' + \mathbf{D}\mathbf{z}\mathbf{z}'), \end{aligned}$$

where $tr(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$, the sum of its diagonal terms. After substituting from (1) for the vectors of totals $\mathbf{y}$ and $\mathbf{z}$ and also using (1) in (3), (4) and (6) it can be shown, with a certain amount of algebraic manipulation, that the differences used for estimating the variance components have expected values

$$E[R(\mu, \alpha, \beta) - R(\mu, \beta)]$$
$$= (n_{..} - k_2)\sigma_\alpha^2 + (k_\beta - k_2)\sigma_{\alpha\beta}^2 + (a - 1)\sigma_e^2 ,$$
$$E[R(\mu, \alpha, \beta, \alpha\beta) - R(\mu, \alpha, \beta)]$$
$$= (n_{..} - k_\beta)\sigma_{\alpha\beta}^2 + (s - a - b + 1)\sigma_e^2 ,$$
$$E[R(0) - R(\mu, \alpha, \beta, \alpha\beta)]$$
$$= (n_{..} - s)\sigma_e^2 .$$

The variances are estimated by equating the right-hand sides of these equations to the calculated values of the differences in the square brackets of the left-hand side. The term k_2 is one of the coefficients used in the random model analysis, namely $k_2 = \sum_{j=1}^{b} [\sum_{i=1}^{a} n_{ij}^2]/n_{.j}$. The coefficient k_β comes from the expected value of $R(\mu, \alpha, \beta)$ and can be expressed in the form

$$k_\beta = 2 \operatorname{tr} (\mathbf{AU} + 2\mathbf{BV} + \mathbf{DW}), \tag{15}$$

where $\mathbf{U}$ is a diagonal matrix of order a with terms $\sum_{j=1}^{b-1} n_{ij}^2$ for $i = 1 \cdots a$, $\mathbf{V}$ is a matrix of order a by $b - 1$ with terms n_{ij}^2 for $i = 1 \cdots a$, and $j = 1 \cdots b - 1$, and W is a diagonal matrix of order $b - 1$ with terms $\sum_{i=1}^{a} n_{ij}^2$ for $j = 1 \cdots b - 1$.

COMPUTING R_β AND k_β

R_β is obtained from computing the following terms:

(i) A square matrix $\mathbf{C}$, of order $(b - 1)$ with terms

$$c_{jj} = n_{.j} - \sum_i n_{ij}^2/n_{i.} , \qquad j = 1, \cdots , b - 1,$$

and

$$c_{jj'} = -\sum_i n_{ij}n_{ij'}/n_{i.}, \qquad j \neq j' = 1, \cdots , b - 1.$$

Computing c_{bj} provides the check that $\sum_{j'=1}^{b} c_{jj'} = 0$, for all j, i.e. row and column totals of the augmented $\mathbf{C}$ are zero.

(ii) A column vector $\mathbf{r}$, of order $b - 1$, whose terms are

$$r_j = x_{.j.} - \sum_i n_{ij}\bar{x}_{i..} , \qquad j = 1, \cdots , b - 1.$$

Computing r_b provides the check that the sum of all the r_j's is zero.

(iii) A column vector, of order $b - 1$, of the estimates of the β's,

$$\hat{\boldsymbol{\beta}} = \mathbf{C}^{-1}\mathbf{r}.$$

(iv) The "inner product" of the terms in the preceding two vectors is R_β:

$$R_\beta = \hat{\boldsymbol{\beta}}'\mathbf{r} = \sum_{j=1}^{b-1} \hat{\boldsymbol{\beta}}_j r_j = \mathbf{r}'\mathbf{c}\mathbf{r}.$$

This is equation (13), the $\mathbf{C}^{-1}$ used here being identical to $\mathbf{D}$.

Expression (15) for k_β is obtained using the elements of $\mathbf{C}^{-1}$, which we shall call $d_{jj'}$, obtained in the calculating of R_β above. The special forms of the matrices involved in (15) enables k_β to be written as

$$k_\beta = \sum_{i=1}^{a} \lambda_i + \sum_{j=1}^{b-1} d_{jj}\left(\sum_{i=1}^{a} f_{i,jj}\right) + 2 \sum_{j' \neq j, =1}^{b-1} \sum^{b-1} d_{jj'}\left(\sum_{i=1}^{a} f_{i,jj'}\right),$$

where

$$\lambda_i = \sum_{j=1}^{b-1} n_{ij}^2/n_{i.}\ ,$$

$$f_{i,jj} = (n_{ij}^2/n_{i.})\,(\lambda_i + n_{i.} - 2n_{ij}),$$

and

$$f_{i,jj'} = (n_{ij}n_{ij'}/n_{i.})\,(\lambda_i - n_{ij} - n_{ij'}).$$

The f's, of which $\frac{1}{2}b(b - 1)$ in number are required, can be calculated for each i and summed, and the expression is well-suited to evaluation on a computer. Desk calculation can be arranged from the same formula when b is small, and the number of α-classes (random) is not too large. Calculating all of the $\frac{1}{2}b(b + 1)$ f's enables the following check to be placed on them:

$$\sum_{j'=1}^{b} \left(\sum_i f_{i,jj'}\right) = -\sum_i n_{ib}^2 n_{ij}/n_{i.}\ . \tag{16}$$

The f-values and $\sum_i \lambda_i$ can be obtained simultaneously with the terms of $\mathbf{C}$ and $\mathbf{r}$; R_β is calculated as $\mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$ and k_β is obtained using the elements of $\mathbf{C}^{-1}$ as above.

EXAMPLE

The calculation of R_β and k_β is demonstrated for the small hypothetical example shown in Table 1.

The steps for obtaining R_β are as follows:

(i) The c-values are

$$c_{11} = 53 - 1^2/10 - 3^2/20 - 12^2/40 - 12^2/50 - 25^2/50 = 33.47,$$

$$c_{22} = 45 - 2^2/10 - 6^2/20 \qquad - 12^2/50 - 25^2/50 = 27.42,$$

TABLE 1

Hypothetical Example

	Number of observations n_{ij}					Mean observed values, $x_{i..}$
		j				
i	1	2	3	4	Totals $n_{i\cdot}$	
1	1	2	3	4	10	200
2	3	6	4	7	20	300
3	12	—	12	16	40	400
4	12	12	13	13	50	500
5	25	25	—	—	50	600
Totals $n_{\cdot j}$	53	45	32	40	$n_{..} = 170$	
Totals of observed values, $x_{\cdot j}$	23000	23000	14000	19000		

$$c_{33} = 32 - 3^2/10 - 4^2/20 - 12^2/40 - 13^2/50 = 23.32,$$

$$c_{44} = 40 - 4^2/10 - 7^2/20 - 16^2/40 - 13^2/50 = 26.17,$$

$$c_{12} = -1(2)/10 - 3(6)/20 - 12(12)/50 - 25(25)/50 = -16.48,$$

$$c_{13} = -1(3)/10 - 3(4)/20 - 12(12)/40 - 12(13)/50 = -\ 7.62,$$

and similarly

$$c_{14} = -9.37, \qquad c_{24} = -\ 6.02,$$

$$c_{23} = -4.92, \qquad c_{34} = -10.78.$$

The check on these values, namely $\sum_{j'=1}^{b} c_{jj'} = 0$, can be seen to hold true; for example

$$c_{11} + c_{12} + c_{13} + c_{14} = 33.47 - 16.48 - 7.62 - 9.37 = 0.$$

(ii) The r-values are obtained using the totals of the observations for the levels of the fixed effects and the means for the levels of the random effects:

$$r_1 = 23000 - 1(200) - 3(300) - 12(400) - 12(500) - 25(600) = -3900,$$

$$r_2 = 23000 - 2(200) - 6(300) \qquad - 12(500) - 25(600) = -200,$$

$$r_3 = 14000 - 3(200) - 4(300) - 12(400) - 13(500) = 900,$$

$$r_4 = 19000 - 4(200) - 7(300) - 16(400) - 13(500) = 3200.$$

A check on these is that their sum is zero.

(iii)

$$\mathbf{C}^{-1} = \mathbf{D} = \begin{bmatrix} 33.47 & -16.48 & -7.62 \\ -16.48 & 27.42 & -4.92 \\ -7.62 & -4.92 & 23.32 \end{bmatrix}^{-1}$$

(iv)

$$R_\beta = \mathbf{r}'\mathbf{D}\mathbf{r}$$

$$= (-3900, -200, 900) \begin{bmatrix} .053824 & .036902 & .025373 \\ .036902 & .063205 & .025393 \\ .025373 & .025393 & .056530 \end{bmatrix} \begin{bmatrix} -3900 \\ -200 \\ 900 \end{bmatrix}$$

$$= 736703.$$

Calculating k_β involves the elements of $\mathbf{D}$ and the f-values. The latter, together with the λ_i-terms are shown in Table 2. The calculations for $i = 1$ are as follows.

$$\lambda_1 = (1^2 + 2^2 + 3^2)/10 = 1.40,$$

$$f_{1,11} = 1^2(1.4 + 10 - 2)/10 = .94,$$

$$f_{1,22} = 2^2(1.4 + 10 - 4)/10 = 2.96,$$

$$f_{1,33} = 3^2(1.4 + 10 - 6)/10 = 4.86,$$

$$f_{1,12} = 1(2)(1.4 - 1 - 2)/10 = -.32,$$

$$f_{1,13} = 1(3)(1.4 - 1 - 3)/10 = -.78.$$

Similarly $f_{1,14} = -1.44$, $f_{1,23} = -2.16$, $f_{1,24} = -3.68$ and $f_{1,34} = -6.72$. The sums of these terms, over all values of i, are shown in Table 2 from which the f-values can be checked by the expression (16); for example

$$\sum_i (f_{i,11} + f_{i,12} + f_{i,13} + f_{i,14})$$

$$= 505.8357 - 360.9718 - 113.1132 - 158.0607$$

$$= -126.31$$

$$= -4^2(1)/10 + 7^2(3)/20 + 16^2(12)/40 + 13^2(12)/50$$

$$= -(1.60 + 7.35 + 76.80 + 40.56).$$

From the c's and the f's k_β is now computed as

$$k_\beta = 45.79 + [505.8357(.053824) + 436.5532(.063205) + 212.4332(.056530)]$$

TABLE 2

Terms Used in Calculating k_β

i	λ_i	$f_{i,11}$	$f_{i,22}$	$f_{i,33}$	$f_{i,44}$	
1	1.40	.9400	2.9600	4.8600	5.4400	
2	3.05	7.6725	19.8900	12.0400	22.1725	
3	7.20	83.5200		83.5200	97.2800	
4	9.14	101.2032	101.2032	112.0132	112.0132	
5	25.00	312.5000	312.5000	—	—	
Total	45.79	505.8357	436.5532	212.4332	236.9057	
i	$-f_{i,12}$	$-f_{i,13}$	$-f_{i,14}$	$-f_{i,23}$	$-f_{i,24}$	$-f_{i,34}$
1	.3200	.7800	1.4400	2.1600	3.6800	6.7200
2	5.3550	2.3700	7.2975	8.3400	20.8950	11.1300
3	—	60.4800	99.8400	—	—	99.8400
4	42.7968	49.4832	49.4832	49.4832	49.4832	56.9868
5	312.5000	—	—	—	—	—
Total	360.9718	113.1132	158.0607	59.9832	74.0582	174.6768

$$- 2[360.9718(.036902) + 113.1132(.025373) + 59.9832(.025393)]$$

$$= 77.16.$$

This procedure for obtaining R_β and k_β may not appear greatly more straight-forward than inverting the matrix required in (8) which in this case is the 8 × 8 matrix

$$\begin{bmatrix} 10 & & & & & 1 & 2 & 3 \\ & 20 & & & & 3 & 6 & 4 \\ & & 40 & & & 12 & 0 & 12 \\ & & & 50 & & 12 & 12 & 13 \\ & & & & 50 & 25 & 25 & 0 \\ 1 & 3 & 12 & 12 & 25 & 53 & & \\ 2 & 6 & 0 & 12 & 25 & & 45 & \\ 3 & 4 & 12 & 13 & 0 & & & 32 \end{bmatrix}.$$

Advantages are apparent however, when one considers the case of a large number of levels of the random classification 500 say, instead of 5. The matrix to be inverted would then be of order 503 while the procedure outlined here would still only require inverting a 3 × 3. The calculation of its terms the c's and of the terms for k_β, the f's, is still lengthy but can be accomplished separately for each i and summation made over i. This can be arranged quite straightforwardly for a desk calculator and is easily organized for an electronic computer such as the IBM 650, for example.

REFERENCES

Eisenhart, C. [1947]. The assumptions underlying the analysis of variance. *Biometrics 3*, 1–21.

Henderson C. R. [1953]. Estimation of variance and covariance components. *Biometrics 9*, 226–52.

Henderson C. R., Kempthorne O., Searle S. R., and Von Krosigk C. M. [1959]. Estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192-218.

Searle S. R. [1958]. Sampling variances of estimates of components of variance. *Ann. Math. Stat. 29*, 167–78.

Searle S. R. and Henderson C. R. [1960]. Judging the effectiveness of age correction factors. *J. Dairy Sci. 43*, 966-74.

OPTIMUM SAMPLE SIZE IN ANIMAL DISEASE CONTROL[1]

A. W. Nordskog[2], H. T. David[3] and H. B. Eisenberg[4]
Iowa State University, Ames, Iowa, U. S. A.

1. INTRODUCTION

The objective of an animal disease control program would either be (*a*) to completely eradicate a disease or alternatively (*b*) to reduce the incidence of a disease to a low level and then to keep it in check. In the case of (*b*) the causative organism might not be completely stamped out.

If a disease has not obtained a strong foothold in a country, then objective (*a*) clearly is justified and perhaps the most drastic methods of control are in order. If, on the otherhand, a disease is widely distributed both geographically and zoologically, then alternative (*b*) might be more realistic. This implies that the objective of a particular control program should take into account the cost of the program relative to the price of the consequence if there were no control.

A case in point is the method of control for Pullorum disease (Salmonella pullorum) in poultry. The active form of the disease causes an enteritis in baby chicks which may result in high death losses. Chicks which recover may be carriers of the disease as adults. Female carriers may then transmit the organism via the egg to their progeny, where the disease again may become epidemic. Other species of domestic birds, and numerous wild species including pheasants, quail, and even foxes, cats, swine, cattle and man, have been reported as potential carriers of Pullorum.

Control of this disease is based on a blood testing technique. Samples of blood, collected from each member of an adult flock, are tested against an antigen of pullorum organisms for agglutinizing antibodies. Positive reactors to the blood test are judged to be disease carriers and therefore are removed from the flock and slaughtered.

The blood agglutination test, carried out in most states for the past three or four decades, has now reduced the disease to one of relatively

[1] Journal Paper No. J-4047 of the Iowa Agricultural and Home Economics Experiment Station, Ames, Iowa. Project No. 1326.

[2] Department of Poultry Husbandry. [3] Department of Statistics. [4] Department of Statistics; now at System Development Corporation.

minor importance. For example, the number of reactors reported by participants in the Iowa Control program was less than .07 of one percent in 1959 while in the Massachusetts program no positive reactors were reported in that year.

When a disease such as Pullorum is widely distributed but at a low incidence on a state-wide or national level, it becomes problematical whether to continue past practices of completely blood testing all birds in a flock, or whether blood testing only a sample, or even no birds at all, of each flock might be more economical.

The present paper investigates the extent to which 100 percent blood testing is justified, assuming random, that is binomial, infestation of adult flocks. Specifically, we attempt to determine the most economical flock proportion to be blood tested in any given year and locale, as a function of certain values and costs and of the prevailing binomial infestation rate.

Economic optimizations in the presence of binomial a priori distributions have been studied previously, as for example in [1], [2], [3] and [4]. However, this prior work is almost entirely concerned with industrial inspection problems. The present paper is intended in part to illustrate the fact that such methodology, initially intended for industrial application, is equally applicable in the biological realm.

We recognize that the assumption of binomial infestation might in certain instances be improved upon, by suitable contagion models for example. However, this assumption seems not unreasonable in the case of Pullorum; in this case, the adult flock is composed of birds which, though able to transmit the disease to their progeny through the egg, have low contagious influence as carriers.

2. ALTERNATIVE FLOCK TESTING POLICIES; VALUES AND COSTS

We consider the following one-parameter (the parameter is n) family of flock testing policies:

> Sample and blood test n birds out of a flock of N. If the sample contains no reactors (reaction assumed equivalent to infection), return the n birds to the flock and blood test no more. If the sample contains one or more reactors, blood test the entire flock; slaughter all discovered reactors.

The relative economic worth of each of these $N + 1$ competing flock testing policies are now evaluated in the light of the following values and costs.

i: blood test cost per bird,
c: carcass value per bird,
$a(k)$: average value of a member of a flock of size N that contains k undiscovered reactors,
$b(k)$: value of a non-reactor from a completely tested flock of size N containing k discovered reactors.

Note that $a(0) = b(0)$ is the value of a member of a flock containing no reactors. Again, the value of a discovered reactor from a completely tested flock is c. Finally, the precise shapes of the function $a(k)$ will depend on the likelihood that the progeny from a mating involving an undiscovered reactor will become acutely infected, thereby precipitating an epidemic in the progeny flock. The shape of the function $b(k)$ will relate largely to loss of good-will. If good-will is not an important factor, it will not be unreasonable to set $b(k) = b(0)$ for all k; indeed, this is the assumption made at the end of Section 5.

3. PRELIMINARY CONSIDERATIONS

Computationally tractable forms of $a(k)$ and $b(k)$ are as follows.

$$\begin{aligned} a(0) &= A,\, a(k) = \alpha \quad \text{for} \quad k \neq 0, \\ b(0) &= A,\, b(k) = \beta \quad \text{for} \quad k \neq 0. \end{aligned} \tag{1}$$

It is shown below that the most profitable n is either 0 or N. If this can, for the moment, be assumed, then the derivation of the most profitable sample size (i.e. the choice between 0 and N) becomes a matter of simple arithmetic, at least for the limiting cases of very small or very large flock size N.

Consider a large number M of birds, grouped into m flocks of $N = M/m$ birds. The case of very small flock size is typified by $N = 1$, while the case of very large flock size is typified by $m = 1$.

For $N = 1$, the average value of the M birds (i.e. one-bird flocks) equals

$$M(1 - \pi)A + M\pi\alpha, \qquad \text{if} \quad n = 0, \tag{2a}$$

$$M(1 - \pi)A + M\pi c - Mi, \qquad \text{if} \quad n = N\,(=1), \tag{2b}$$

so that 100 percent testing will be more (less) profitable than no testing according to whether

$$\pi(c - \alpha) > (<)\, i. \tag{3}$$

For $m = 1$, the average value of the M birds (i.e. one flock of M birds) equals

$$M(1 - \pi)\alpha + M\pi\alpha = M\alpha, \qquad \text{if} \quad n = 0, \tag{4a}$$

$$M(1 - \pi)\beta + M\pi c - Mi, \qquad \text{if} \quad n = N\ (=M), \tag{4b}$$

so that 100 percent testing will be more (less) profitable than no testing according to whether

$$\beta - \alpha > (<)\ \pi(\beta - c) + i. \tag{5}$$

Expressions (2*a*) and (2*b*) arise from the fact that, under both zero and complete testing, a proportion $(1 - \pi)$ of the M one-bird flocks will be disease-free, contributing an amount $M(1 - \pi)\ A$ to average value. In addition, there will be a value contribution from the $M\pi$ one-bird flocks containing a reactor; this contribution will amount to $M\pi\alpha$ in the case of no testing and to $M\pi c$ in the case of 100 percent testing. Finally, there is the testing cost Mi that is incurred under 100 percent testing.

Expressions (4a) and (4b) arise from the fact that a very large flock will contain an approximate proportion π of reactors; hence, under 100 percent testing, there will be approximately $M\pi$ birds contributing carcass value c to the flock, and approximately $M(1 - \pi)$ birds contributing value β. Similarly, if M is large enough to make $M\pi$ large, the flock of M birds will contain at least one reactor with probability essentially 1, leading to an average per-bird value of α in the absence of testing.

The criteria represented by (3) and (5) constitute an almost adequate solution for the case of the value assumptions given in (1). The further computations of Section 4 will serve only to validate the assertion that the most profitable sample size must either be zero or N, and will lead, as well, to the analogues of (3) and (5) for flocks of intermediate size.

4. THE COMPUTATION OF THE MOST PROFITABLE SAMPLE SIZE FOR THE CASE OF CONSTANT VALUE DIFFERENCE $b(k) - a(k)$

This is the case typified by form (1) of the functions $a(k)$ and $b(k)$. The computations proceed as follows. Let

$$\phi_k = \text{the binomial probability of } k \text{ reactors in a flock of size } N = \binom{N}{k} \pi^k (1 - \pi)^{N-k}, \tag{6}$$

and let

$h_{n,k}$ = the hypergeometric probability of obtaining n nonreactors when drawing a random sample of size n from a flock of N birds containing k reactors and

$$(N - k) \text{ non-reactors} = \binom{N-k}{n} \Big/ \binom{N}{n}. \tag{7}$$

Then for any particular one of the $(N + 1)$ alternative policies presented in Section 2, say the policy corresponding to sample size n,

$$\begin{aligned}&\Pr \{\text{number of reactors in the flock} = k;\ \text{number of re-}\\ &\quad \text{actors in the sample} = 0\} = \phi_k h_{n,k}\,,\end{aligned} \tag{8}$$

and the "profit" ensuing if the number of reactors in the flock equals k and the number of reactors in the sample equals 0 is

$$Na(k) - ni. \tag{9}$$

Hence the contribution to expected profit of the events involving no reactors in the sample equals the sum of the products of (8) and (9):

$$\sum_{k=0}^{N-n} [Na(k) - ni]\cdot\phi_k\cdot h_{n,k}\,. \tag{10}$$

Again, for the same sample size n,

$$\begin{aligned}&\Pr \{\text{numbers of reactors in the flock} = k;\ \text{number of re-}\\ &\quad \text{actors in the sample} > 0\} = \phi_k\cdot(1 - h_{n,k}),\end{aligned} \tag{11}$$

and the "profit" ensuing if the number of reactors in the flock equals k and the number of reactors in the sample exceeds 0 is

$$(N - k)\cdot b(k) + kc - Ni. \tag{12}$$

Hence the contribution to expected profit of the events involving one or more reactors in the sample equals the sum of the products of (11) and (12):

$$\sum_{k=0}^{N} ((N - k)\cdot b(k) + kc - Ni)\cdot\phi_k\cdot(1 - h_{n,k}). \tag{13}$$

Total expected profit will equal the sum of (10) and (13), which, neglecting terms not involving n and using the fact that $h_{n,k} = 0$ for $k \geq N - n + 1$, can be written

$$\sum_{k=0}^{N-n} \{N\cdot[a(k) - b(k)] + k\cdot[b(k) - c] + (N - n)\cdot i\}\cdot\phi_k\cdot h_{n,k}\,. \tag{14}$$

Expression (14) is further reducible to

$$(1 - \pi)^n\cdot\{N\cdot E[a(k) - b(k)] + E[k\cdot(b(k) - c)] + (N - n)\cdot i\}, \tag{15}$$

where the expectation $E[\ ]$ is with respect to the chance variable k having a binomial distribution with parameters $(N - n)$ and π. This type of expectation arises from the fact that

$$\phi_k h_{n,k} = (1 - \pi)^n \left[\binom{N - n}{k} \pi^k (1 - \pi)^{N-n-k} \right].$$

Further reduction leads, except for the additive term $(1-\pi)^N \cdot N(\beta-\alpha)$, to the expression

$$(1 - \pi)^n (T - nS) \equiv P(n), \tag{16}$$

where

$$T = N[\alpha - \beta + \pi(\beta - c) + i], \tag{17}$$

$$S = \pi(\beta - c) + i > 0. \tag{18}$$

Differencing $P(n)$ now yields

$$\Delta(n) = P(n + 1) - P(n) = (1 - \pi)^n[(n\pi - 1)S - \pi(T - S)],$$

which shows that $\Delta(n)$ is negative for $n < (T/S) + (1/\pi) - 1$, and is positive for $n > (T/S) + (1/\pi) - 1$. This implies that no $P(n)$ for $0 < n < N$ can be larger than the larger of $P(0)$ and $P(N)$, which in turn implies that the most "profitable" sample size n either is 0 or N. Establishing this fact (which is reminiscent of conclusions reached in [1] and [2]) was one of the two objectives set for this section in the last paragraph of Section 3.

The second objective set for this section was to derive a condition analogous to (3) and (5) for determining the relative profitabilities of the two sample sizes 0 and N for intermediate flock size N. But this now is simply a matter of comparing $P(0)$ and $P(N)$, where $P(n)$ is given by (16). This yields the conclusion that 100 percent testing is more (less) profitable than no testing at all according to whether

$$(\beta - \alpha)[1 - (1 - \pi)^N] > (<)\, \pi(\beta - c) + i. \tag{19}$$

We note that (19) does indeed specialize to (3) and (5) for N equal, respectively, to 1 and to ∞.

Although details are outside the scope of this paper, it may be of interest to point out that condition (3) arises naturally in the computation of the sequential Bayes test policy for the value assumptions (1). Consider the much (though no doubt impractically) enlarged set of flock testing policies consisting of all sequential plans with terminal acts A and R:

A : Stop testing; collect c for every reactor culled out so far; return non-reactors to the flock; eventually collect α or β per bird of the unculled portion of the flock, depending on whether or not this portion contains at least one reactor, unless the entire flock is reactor-free, in which case collect A.

R : Test 100 percent; collect c for reactors, and β for non-reactors, unless the entire flock is reactor-free, in which case collect A.

A straightforward application of the methodology given in [2] then shows that 100 percent testing is the most economic of the policies in this enlarged set if $\pi(c - \alpha) > i$. However, the complementary prescription for no testing if $\pi(c - \alpha) < i$ does *not* apply in this case.

5. THE COMPUTATION OF THE MOST PROFITABLE SAMPLE SIZE FOR THE CASE OF LINEARLY INCREASING VALUE DIFFERENCE $b(k) - a(k)$.

A value assumption alternative to (1) is

$$\begin{aligned} a(k) &= A - \frac{k}{N}(A - \gamma), \qquad 0 \leq k \leq N, \\ b(k) &= A - \frac{k}{N}(A - \delta), \qquad 0 \leq k \leq N. \end{aligned} \tag{20}$$

Using (15) [which was derived without reference to any specific form of $a(k)$ and $b(k)$], expected profit now becomes, except for additive constants not involving n,

$$P(n) = (1 - \pi)^n \left(\frac{N - n}{N}\right)(T + nS), \tag{21}$$

where

$$T = (A - \delta)(\pi - \pi^2)(N - 1) + N[\pi(\gamma - c) + i], \tag{22}$$

$$S = (A - \delta)\pi^2 \geq 0. \tag{23}$$

The function $P(n)$ given by (21) is best described in terms of the following five parametric cases.

Case I: $S > 0$, $T \geq 0$. As n increases from $-\infty$, $P(n)$ rise steadily from $-\infty$, crosses the n-axis at $n = -T/S$, equals T at $n = 0$, turns downward somewhere between $n = -T/S$ and $n = N$, crosses the n-axis once more at $n = N$, turns upward somewhere beyond $n =$ N, and approaches the n-axis asymptotically from below as n approaches $+\infty$.

Case II: $S > 0$, $0 > T > -NS$. As n increases from $-\infty$, $P(n)$ rises steadily from $-\infty$, equals T at $n = 0$, crosses the n-axis at $n = -T/S$, turns downward somewhere between $n = -T/S$ and $n = N$, crosses the n-axis once more at $n = N$, turns upward somewhere beyond $n = N$, and approaches the n-axis asymptotically from below as n approaches $+\infty$.

Case III: $S > 0$, $T \leq -NS$. As n increases from $-\infty$, $P(n)$ rises steadily from $-\infty$, equals T at $n = 0$, crosses the n-axis at $n = N$, turns downward somewhere between $n = N$ and $n = -T/S$, crosses the n-axis once more at $n = -T/S$, turns upward somewhere beyond $n = -T/S$, and approaches the n-axis asymptotically from below as n approaches $+\infty$.

Case IV $S = 0$; $T > 0$. As n increases from $-\infty$, $P(n)$ decreases steadily from $+\infty$, equals T at $n = 0$, crosses the n-axis at $n = N$, turns upward somewhere beyond $n = N$, and approaches the n-axis asymptotically from below as n approaches $+\infty$.

Case V: $S = 0$, $T \leq 0$. As n increases from $-\infty$, $P(n)$ increases steadily from $-\infty$, equals T at $n = 0$, crosses the n-axis at $n = N$, turns downward somewhere beyond $n = N$, and approaches the n-axis asymptotically from above as n approaches $+\infty$.

Cases IV and V are easily summarized as follows: If $A = \delta$ (corresponding, as indicated at the end of Section 2, to the absence of the good-will factor), then the only policies in contention are 100 percent testing ($n = N$) and no testing ($n = 0$), and 100 percent testing will be more (less) profitable than no testing according to whether

$$\pi(c - \gamma) > (<)\, i. \tag{24}$$

It seems of interest to note the resemblance of (24) and (3).

Cases III and V are summarized by: If $T + NS \leq 0$, it is most economical to test 100 percent.

For Case II, the most economical sample size is the n between $-T/S$ and N at which $P(n)$ turns downward. This case is of interest since it establishes the possibility of an optimum sample size other than 0 or N. This possibility has already been noted in [3].

For Case I, the most economical sample size is either $n = 0$ or the n at which $P(n)$ turns downward, depending on the relative magnitudes at these two sample sizes. Note that, for $T = 0$, the n at which $P(n)$ turns downward must be greater than zero, so that, as in Case II, the most economical sample size will be other than 0 or N.

6. EXAMPLE AND CONCLUSIONS.

Consider the case when good-will is not a factor, i.e. the case of constant $b(k)$. In this case both the formulation of Section 4 and that of Section 5 imply that only the two policies of zero and 100 percent testing are in contention, the choice between these depending on the direction of a simple inequality.

Defining the *critical testing cost* i_c to be the testing cost for which zero and 100 percent testing are equally economical, it seems of interest to compute i_c for both formulations, using comparable value figures.

A typical value for a non-reactor in a fully tested flock is \$3, a typical average value of a member of an untested flock is \$2.40, and typical values for c, π and N are \$.50, 10^{-5} and 500.

Computing i_c in the spirit of Section 4, we therefore set $A = \beta = \$3$ and $\alpha = \$2.40$. Replacing $1 - (1 - \pi)^N$ by $N\pi$ (allowable since $N\pi$ is small), criterion (19) then becomes $\$3 \times 10^{-3} > (<) \$2.5 \times 10^{-5} + i$, which means that $i_c = \$0.003$, i.e. that no testing is most profitable unless the cost of testing falls below 3 mills per bird.

Computing i_c in the spirit of Section 5, we set $A = \delta = \$3$. In addition, we interpret \$2.40 to be the value of the linear function $a(k)$ evaluated at $k = N\pi$, the expected number of reactors in the flock. This implies a per-bird disaster value of $\gamma = -\$6 \times 10^4$ for an unsuspected 100 percent infected flock. Criterion (24) then becomes $\$0.6 > (<) i$, which means that $i_c = \$0.60$, i.e. that 100 percent testing is most profitable unless the cost of testing rises above 60¢ per bird.

In practice, the cost of testing is approximately seven cents per bird. Since this cost is well bracketed by the critical costs 0.3¢ and 60¢ derived above, we learn that the shape of the value function $a(k)$ [and of course also that of $b(k)$] must be determined rather accurately if the methodology presented here is to be applied.

REFERENCES

[1]. Barnard, G. A. [1954]. Sampling inspection and statistical decisions. *JRSS (B), 16,* 151–72.

[2]. Eisenberg, H. B. [1959]. Bayesian lot-by-lot sampling inspection. *M. S. Thesis* Iowa State University, Ames, Iowa.

[3]. Eisenberg, H. B. [1960]. Bayesian sampling inspection for binomial a priori distributions and quadratic loss functions. *Unpublished.*

[4]. Hald, A. [1960]. The compound hypergeometric distribution and a system of single sampling inspection plans based on prior distributions and costs. *Technometrics 2,* 275–340.

NUMERICAL ASPECTS OF THE REGRESSION OF OFFSPRING ON PARENT[1]

H. E. McKean and B. B. Bohren
Population Genetics Institute, Purdue University
Lafayette, Indiana, U. S. A.

INTRODUCTION

In an earlier paper (Bohren, McKean, and Yamada, [1961]) three currently employed techniques for estimating the regression of offspring on parent, and thereby heritability in the narrow sense, were compared and contrasted with respect to their efficiencies of estimation. The general conclusion, based on theoretical considerations and an empirical study of five generations of a closed poultry flock (Yamada, Bohren, and Crittenden, [1957]), was that under the circumstances considered, the method of regression of offspring means on parent's record (method 1) was inferior to the method of regression of individual offspring on parent (method 2) (with the parent's record repeated once for each of its offspring) and to (method 3) the Kempthorne-Tandon technique (Kempthorne and Tandon, [1953]).

The success of the Kempthorne-Tandon technique depends upon knowledge of a parameter ρ, the correlation between deviations of two offspring of the same parent from the predicted breeding value of the parent, and its expected superiority over the second method depends upon the magnitude of ρ. Usually ρ is guessed in the light of prior knowledge, and weights are assigned to the families according to the guessed value of ρ. In the first paper it was shown that, under the assumption of all genetic variance being additive, $\rho \leq .067$ or $\leq .079$, depending upon whether the mating structure is random or hierarchal.

The results obtained in the previous paper specifically depended upon the particular distribution of family sizes encountered in the five analyses, and upon the accuracy of the estimated values of $T = \rho/(1-\rho)$. The purposes of this paper are to consider the efficiency loss incurred by mis-guessing ρ in the Kempthorne-Tandon technique, and to investigate the factors involved in the relative efficiency of the other two methods.

THEORY

We consider a breeding experiment in which s sires are selected

[1] Journal Paper Number 1688, Purdue University Agricultural Experiment Station.

from the population, sire i being mated to a random sample of d_i distinct dams. The mating structure is then hierarchal, with the progeny of sire i having phenotypic values given by the model

$$Y_{ijk} = \mu_i + \beta(X_{ij} - \mu) + e'_{ijk} , \tag{1}$$

which is equation (11) of the previous paper.

It has been pointed out that the three methods under consideration are merely special cases of the general unbiased weighted regression estimation procedure. The difference between the methods involves only the weights applied to each progeny-group deviation: (1) $w_{ij} = 1$ for the progeny means on parents technique, (2) $w_{ij} = n_{ij}$ for the repeated parents technique, and (3) $w_{ij} = n_{ij}/(1 + n_{ij}\tau)$ for the Kempthorne-Tandon technique, where τ is a guessed value of T. If $\tau = T$, the third technique is the minimum variance technique, whereas, if $\tau = 0$, the third technique reduces to the second. Furthermore, when the family sizes are equal ($n_{i1} = n_{i2} = \cdots$, for all i), all three methods are identical in efficiency. Since most experimental data will involve unequal family sizes, interest will center on considering this situation.

The question, "When may I use method 1 with little loss in efficiency?", is a pertinent one. First, let us examine the optimum choice of weights for which the variance of the estimate of β will be minimized. These optimal weights w_{ij}^* (say), where $w_{ij}^* = n_{ij}/(1 + n_{ij}T)$, will be approximately equal (hence method 1 appropriate), for $T > 0$, under one or more of three distinct sets of circumstances:

1) All n_{ij} are large. This follows immediately from

$$\lim_{n_{ij}\to\infty} \left[\frac{n_{ij}}{1 + n_{ij}T}\right] = \frac{1}{T}.$$

2) S_n^2 , the variance of the family sizes, is zero or very small.

3) T is large. This follows from the fact that

$$\lim_{\substack{T\to\infty \\ (\text{or } \rho\to 1)}} \frac{w_{ij}^*}{w_{i'j'}^*} = 1,$$

independent of n_{ij} and $n_{i'j'}$; thus for large T, $w_{ij}^* \doteq w_{i'j'}^*$. In view of this, method 1 may also be considered as a special case of method 3 where τ, the guessed value of T, is allowed to approach ∞.

It is of interest, therefore, to determine a readily accessible statistic which depends upon $\bar{n}$ (the average dam family size), S_n^2 , and T, upon which a decision to use or not use method 1 as opposed to method 2 may be based.

It is easy to show that the coefficient of variation among the optimal

$$\exp \{\lambda[(q - pz)^{-k} - 1]\}, \ \lambda > 0, \ k > 0, \ p > 0, \ q = 1 + p. \qquad (1)$$

It is believed that most of the deviation of the distribution of the survivors from the simple distributions mentioned above stems from (i) the complex structure of survivors within an egg mass and (ii) the movement of survivors from place to place. Since the Negative Binomial distribution has been found to be very useful in fitting data involving this type of heterogeneity, it is reasonable to suppose that the Poisson Pascal will give a better description of the population of the survivors in a field.

The Poisson Pascal distribution can also be looked upon as the result of compounding a Pascal distribution with p.g.f. $(q - pz)^{-k_1}$ by taking k_1 to behave as kx where k is a positive constant and x a Poisson random variable with p.g.f. $\exp \{\lambda(z - 1)\}$.

3. SOME PROPERTIES OF THE POISSON PASCAL DISTRIBUTION

The limiting forms that the Poisson Pascal distribution takes as the parameters take on extreme values are given in Table 1. It is to be noted that the Neyman Type A and the Pascal distributions are among the limiting forms.

TABLE 1

SOME LIMITING FORMS OF THE POISSON PASCAL DISTRIBUTION

No.	Limits Taken	Name and p.g.f. of the limit
1	$k \to \infty, p \to 0$ $pk = \lambda_1$	Neyman Type A, $\exp \{\lambda[\exp (\lambda_1(z - 1) - 1]\}$
2	$k \to 0, \lambda \to \infty$ $\lambda k = k_1$	Negative Binomial, $(q - pz)^{-k_1}$
3	$p \to 0, \lambda \to \infty$ $\lambda kp = \lambda_1$	Poisson, $\exp \{\lambda_1(z - 1)\}$

The flexibility of the Poisson Pascal was compared quantitatively with that of the Neyman Types A, B, C, Pascal, and Poisson Binomial by evaluating the relative skewness and kurtosis of each for fixed mean k_1p_1 and variance $k_1p_1(1 + p_1)$ using the indices $\kappa_{[3]}/k_1p_1^3$ and $\kappa_{[4]}/k_1p_1^4$ respectively along the lines of Anscombe [1]. The range of numerical values of these indices for each of the foregoing distributions is shown in Table 2. It is apparent that the Poisson Pascal covers the entire range of distributions from Neyman Type A to Pascal with respect to skewness and kurtosis. Also, the ranges of these ratios for

TABLE 2

Comparison of Skewness and Kurtosis of Certain Distributions

No.	Name	Range of Skewness	Range of Kurtosis
1	Neyman Type A	1	1
2	Neyman Type B	9/8	27/20
3	Neyman Type C	6/5	8/5
4	Pascal	2	6
5	Poisson Pascal	(1,2)	(1,6)
6	Poisson Binomial	(0,1)	(0,1)

the Poisson Pascal and the Poisson Binomial are disjoint. Since their p.g.f.s have the common form

$$\exp \{\lambda[(q - pz)^{-k} - 1]\}, \tag{2}$$

we observe that the distribution with (2) for p.g.f. wherein $\lambda > 0$, $q = 1 + p$, $p > 0$ when $k > 0$ and $-1 < p < 0$ when k is a negative integer. covers a very wide range of distributions. As an aid in computing the values of the ratios for the samples to obtain an idea as to how close the sample distribution is to the various distributions mentioned in Table 2, formulae to compute the factorial cummulants $\kappa_{[i]}$ using sample moments are given in Appendix A.

Since the first two frequencies were large in the sets of data to which this (and similar) distributions were fitted, the ratio of the first two frequencies were compared for some of these distributions. It can be easily shown that the value of this ratio for the Poisson Pascal distribution lies between the ratios for the Neyman Type A and the Pascal distributions. The ratios are given for brevity in Table 3.

TABLE 3

Comparison of the Ratio of the First two Frequencies with Mean and Variance Fixed as k_1P_1 and $k_1P_1(1 + p_1)$

No.	Distribution	Ratio of Frequencies
1	Neyman Type A	$k_1p_1 \exp(-p_1)$
2	Neyman Type B	$2k_1 \{1 - \exp(-3p_1)(1 + 3p_1)\}/(p_1q_1)$
3	Neyman Type C	$3k_1\{2p_1 \exp(-4p_1) + 2[-2 \exp(-4\mathrm{p}_1)$ $+ 64p_1^2 + 4p_1]\}/4p_1^3$
4	Pascal	$k_1p_1(1 + p_1)^{-1}$
5	Poisson Pascal	$k_1p_1(1 + p_1)/(k + 1)^{-k-1}$

4. FITTING POISSON PASCAL AND EFFICIENCY OF METHODS OF ESTIMATION

Since obtaining maximum likelihood estimates is very cumbersome, *ad hoc* methods were used to estimate the parameters. When the mean and the variance were large, use was made of the method of the first three moments. When they were moderate and the proportion of the zero frequency large, use was made of the method of the first two moments and the proportion of the zero frequency. When the first two frequencies were large in comparison with the remaining, the method of the first two moments and the ratio of the first two frequencies was used in estimation. The equations for estimation are given in Appencix C. The fit of this distribution to the data of Beall and Rescia [2] are given in Tables 4 and 5. For the sake of reference, the fit of a generalization of the Neyman Type A as given by the authors of these data are also given alongside. The relatively good fit of the Poisson Pascal is apparent.

For obtaining a comparison of the various methods of estimation, it was decided to compute the efficiency function

'TABLE 4

Fit of the Observed Frequency of Lespedeza Capitata, Table V of [2]

Plants	Observed Frequency	Expected frequency due to Poisson Pascal (Method of Moments)	Expected frequency as in [2]
0	7178	7185.0	7217.6
1	286	276.0	218.6
2	93	94.5	105.5
3	40	41.5	50.9
4	24	20.2	24.5
5	7	10.4	11.8
6	5	5.6	5.7
7	1	3.1	2.8
8	2	1.7	1.3
9	1	1.0	.6
10	2	.6	.3
11+	1	.3	.4
χ^2	—	9.58	42.97
Degrees of Freedom	—	8	9

$E = 1/(\text{Generalized variance} \times \text{Information determinant})$ (cf. Cramer [4], pp. 489–497). (3)

If we denote the parameter vector (λ, p, k) by $(\lambda_1, \lambda_2, \lambda_3)$ and the set of statistics used by (t_1, t_2, t_3), we get the expression for the generalized variance of the estimates as

TABLE 5

Fit of Observed Frequency of Leptinotarsa Decemlineata, Table III of [2]

Insects	Observed	Expected frequency due to Poisson Pascal, (method of two moments and first frequency	Expected frequency as in [2]
0	33	33.0	39.5
1	12	9.8	6.0
2	5	7.4	4.9
3	6	5.5	3.4
4	5	4.0	3.2
5	0	2.9	2.5
6	2	2.1	2.0
7	2	1.5	1.6
8	2	1.1	1.3
9	0	.8	1.0
10	1	.6	.8
11+	2	1.3	3.3
χ^2	—	6.88	13.75
Degrees of Freedom	—	8	9

$$G = \mid V(t_1,t_2,t_3) \mid \Big/ \left| \frac{\partial(\tau_1, \tau_2, \tau_3)}{\partial(\lambda_1, y_2, \lambda_3)} \right|^2, \tag{4}$$

where τ_1, τ_2, τ_3 are the functions of λ_1, λ_2 and λ_3 estimated consistently by t_1, t_2, t_3. A proof of this is given in Appendix B. Since evaluating the covariance matrix $V(t_1, t_2, t_3)$ and the derivatives of t_1, t_2 and t_3 follows from the regular statistical techniques, no elaboration need be made here. A formula for the information determinant (cf. Shenton [10]) is

$$n^3 I = \begin{vmatrix} \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_1} \frac{\partial P_x}{\partial \lambda_1} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_1} \frac{\partial P_x}{\partial \lambda_2} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_1} \frac{\partial P_x}{\partial \lambda_3} \\ \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_2} \frac{\partial P_x}{\partial \lambda_1} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_2} \frac{\partial P_x}{\partial \lambda_2} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_2} \frac{\partial P_x}{\partial \lambda_3} \\ \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_3} \frac{\partial P_x}{\partial \lambda_1} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_3} \frac{\partial P_x}{\partial \lambda_2} & \sum_x \frac{1}{P_x} \frac{\partial P_x}{\partial \lambda_3} \frac{\partial P_x}{\partial \lambda_3} \end{vmatrix}. \tag{5}$$

The principal problems in evaluating I for a value of (λ, p, k) therefore are (*i*) to evaluate the various P_x and the derivatives of P_x and (*ii*) to determine the number of terms to be used in summing the infinite series. To obtain P_x, let

$$g(z) = \exp \{\lambda[(q - pz)^{-k} - 1]\} \tag{6}$$

and

$$h(z) = (q - pz)^{-k}.$$

By differentiating (6) successively we get

$$g'(z) = \lambda g(z) h'(z), \tag{7}$$

and

$$g^{(x+1)}(z) = \lambda \sum_{i=0}^{x} \binom{x+1}{i} g^{(i)}(z) h^{(x+1-i)}(z). \tag{8}$$

Set $z = 0$ in equations (6), (7) and (8) and observe that $g^{(x)}(0) = x!\, P_x$ and $h^{(x)}(0) = x!\, \pi_x$ where

$$\pi_x = \frac{(k + x - 1)!}{(k - 1)!\, x!} p^x q^{-k-x}, \tag{9}$$

is the probability of x in the Pascal distribution with $h(z)$ as p.g.f. Then we have the recurrence formulae

$$P_0 = \exp. \{\lambda[(q)^{-k} - 1]\} \tag{10}$$

and

$$P_{x+1} = \frac{\lambda}{x + 1} \left\{ \sum_{i=0}^{x} (x + 1 - i) \pi_{x+1} - i\, P_i \right\}, \tag{11}$$

which can be repeatedly used to evaluate any P_x.

The various derivative are given by

$$\frac{\partial P_x}{\partial \lambda_1} = \frac{\partial P_x}{\partial \lambda} = \left\{ \frac{\partial^x}{\partial z^x} \left[\frac{\partial g(z)}{\partial \lambda} \right] \right\}_{z=0} = \frac{1}{x!} \left\{ \frac{\partial^x}{\partial z^x} [g(z)(h(z) - 1)] \right\}_{z=0}$$

$$= \frac{q}{\lambda k p} (x + 1) P_{x+1} - P_r \left(1 + \frac{r}{\lambda k} \right), \tag{12}$$

$$\frac{\partial P_x}{\partial \lambda_2} = \frac{\partial P_x}{\partial p} = \frac{1}{x!}\left\{\frac{\partial^x}{\partial z^x}\left[\frac{\partial}{\partial P}\, g(z)\right]\right\}_{z=0} = \frac{1}{x!}\left\{\frac{\partial^x}{\partial z^x}\left[\frac{z}{p}\, g'(z)\right]\right\}_{z=0}$$

$$= \frac{1}{p}\,\{xP_x - (x+1)P_{x+1}\} \qquad (13)$$

and

$$\frac{\partial P_x}{\partial \lambda_3} = \frac{\partial P_x}{\partial k} = \frac{1}{x!}\left\{\frac{\partial^x}{\partial z^x}\left[\frac{\partial}{\partial r}\, g(z)\right]\right\}_{z=0}$$

$$= \frac{1}{x!}\left\{\frac{\partial^x}{\partial z^x}\,[-\lambda g(z)h(z)\log(q - pz)]\right\}_{z=0}$$

$$= \frac{1}{kp}\sum_{i=1}^{x}(p/q)^i\,\frac{1}{i}\,\{q(x-i+1)P_{x-i+1} - p(x-i)P_{x-i}\}$$

$$- \frac{1}{kp}\,(\log q)\{q(x+1)P_{x+1} - pxP_x\} \qquad (14)$$

for all x.

TABLE 6

EFFICIENCY OF THE METHOD OF THE FIRST THREE MOMENTS FOR THE POISSON PASCAL

				k		
λ	p	.1	.3	.5	1.0	2.0
.1	.1	.84	.82	.82	.81	.76
.1	.3	.59	.58	.58	.54	.47
.1	.5	.45	.44	.43	.40	.33
.1	1.0	.26	.25	.24	.22	.18
.1	2.0	.13	.12	.12	.12	.13
.5	.1	.90	.81	.82	.77	.67
.5	.3	.59	.58	.56	.49	.35
.5	.5	.46	.44	.41	.34	.22
.5	1.0	.26	.25	.23	.18	.10
.5	2.0	.13	.13	.12	—	—
1.0	.1	.81	.83	.82	.75	.63
1.0	.3	.59	.59	.55	.46	.28
1.0	.5	.46	.44	.40	.31	.15
1.0	1.0	.27	.25	.23	.15	.05
1.0	—	—	—	—	—	—
5.0	.1	.94*	.99*	.78*	.58*	.21*
5.0	.3	.62*	.56*	.41*	.11*	.01*
5.0	.5	.48*	.38*	.20*	.04*	.00*
5.0	1.0	.29*	.17*	.06*	.02*	.00*
5.0	—	—	—	—	—	—

To determine the number of terms we use the following rule:

Let $T_{ij}(n)$ denote the $(n + 1)$th term in the series involved in the (i, j)th term of matrix (5). Let $S_{ij}(n) = \sum_{r=0}^{n} T_{ij}(r)$. Compute $S_{ij}(n)$ and $\sum_{ij} T^2_{ij}(n)/S^2_{ij}(n)$ for $n = 1, 2, \cdots$ *et cetera* successively till a value of n is reached for which

$$\sum_{ij} T^2_{ij}(n)/S^2_{ij}(n) < 10^{-8}. \tag{15}$$

It is clear from (15) that $| T_{ij}(n) |/| S_{ij}(n) | < 10^{-4}$ for each i and j. If the series converge faster than a geometric series with common ratio less than 0.9 and this convergence starts before the value of n is reached for which (15) is satisfied, the calculated efficiency will be correct to three significant figures. If the significant figures do not cancel out, this should yield the efficiencies computed therefrom, correct to three decimal places. When the inequality (15) was not satisfied for values of $n \leq 20$, the partial sum of the first twenty terms was taken as the value of the series since evaluating the terms of the series

TABLE 7

EFFICIENCY OF THE METHOD OF THE FIRST TWO MOMENTS AND THE FIRST FREQUENCY FOR THE POISSON PASCAL DISTRIBUTION AT λ = 0.1

				k		
λ	p	.1	.3	.5	1.0	2.0
.1	.1	.99	.98	.98	.98	.98
.1	.3	.93	.94	.95	.94	.93
.1	.5	.90	.90	.91	.91	.89
.1	1.0	.82	.83	.84	.84	.83
.1	2.0	.74	.76	.78	.84	.78
.5	.1	1.00	1.00	.98	.96	.95
.5	.3	.93	.94	.93	.92	.87
.5	.5	.91	.91	.90	.89	.82
.5	1.0	.83	.85	.85	.82	.73
.5	2.0	.75	.78	.81	—	—
1.0	.1	1.00	.99	.97	.96	.94
1.0	.3	.94	.96	.94	.93	.87
1.0	.5	.92	.92	.91	.98	.81
1.0	1.0	.84	.85	.86	.82	.72
1.0	2.0	.88*	.87*	.87*	.87*	.83*
5.0	.1	—	—	.99*	.89*	.98*
5.0	.3	.97*	.99*	1.00*	1.00*	.98*
5.0	.5	.95*	.97*	.98*	.99*	.96*
5.0	1.0	.90*	.96*	.97*	—	—
5.0	2.0	.89*	—	—	—	—

for n larger than 20 is very time consuming. The efficiency when n was restricted to 20 is marked with an asterisk to indicate that they are less likely to be correct to three decimal places. When the efficiency so computed was larger than one (due to the inaccuracy in computing the information determinant), the corresponding cell in the efficiency table is left blank.

The efficiency of the method of moments is given for certain values of (λ, p, k) in Table 6. The efficiency of the method of the first two moments and the first frequency is given in Table 7 and that of the method of the first two moments and the ratio of the first two frequencies in Table 8 for the same values of (λ, p, k).

It is apparent that the method of the first two moments and the ratio of the first two frequencies has high efficiency and is superior to the other two methods when λ is small (and consequently the first two counts account for a large proportion of the observed frequencies). Also the method of the first two moments and the first frequency is highly efficient when λ is moderately large. When λ, p or k approaches

TABLE 8

EFFICIENCY OF THE METHOD OF THE FIRST TWO MOMENTS AND THE RATIO OF THE FIRST TWO FREQUENCIES FOR THE POISSON PASCAL AT $\lambda = 0.1$

				k		
λ	p	.1	.3	.5	1.0	2.0
.1	.1	1.00	1.00	1.00	1.00	1.00
.1	.3	.98	.99	.99	.99	.98
.1	.5	.99	.98	.98	.97	.94
.1	1.0	.94	.94	.93	.91	.83
.1	2.0	.87	.88	.89	.91	.99
.5	.1	1.00	.99	1.00*	1.00	.99
.5	.3	.98	1.00	.99	.98	.96
.5	1.0	.99	.98	.97	.96	.92
.5	2.0	.94	.94	.93	.89*	.80*
1.0	.1	.98	1.00	1.00	1.00	.99
1.0	.3	.99	1.00	.99	.99	.95
1.0	.5	1.00	.99	.98	.96	.91
1.0	1.0	.95	.94	.93	.88	.79
1.0	2.0	.88*	.89*	.91*	—	—
5.0	.1	—	—	.99*	.95*	.91*
5.0	.3	1.00	.99*	.98*	.89*	.76*
5.0	.5	1.00	.96*	.91*	.80*	.65*
5.0	1.0	.95*	.88*	.79*	.84*	1.00*
5.0	2.0	.90*	.98*	—	—	—

infinity, it can be shown by calculus that the method of moments is much superior to the other two but since the efficiency of each of these methods tends to zero for such values, this has little significance.

5. CONCLUSIONS

On the basis of the properties discussed in Section 3 and the fitting in Section 4, we observe that the Poisson Pascal distribution acts as a bridge between the Neyman Type A and the Negative Binomial distributions and may be used with advantage when the latter distributions are inadequate to represent the population accurately.

From the tables of efficiency, it is clear that in the region of tabulations, at least one of the *ad hoc* methods of estimation suggested above has high efficiency. It is believed that in practice, (λ, p, k) will not be far beyond the region of tabulation and that one of these methods can be used without too much loss of information. Techniques for choosing one of the many *ad hoc* methods on the basis of the sample will be discussed in a future paper.

REFERENCES

1. Anscombe, F. J. [1950]. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika 37*, 358–82.
2. Beal, G. and Rescia, R. [1953]. A generalization of Neyman's contagious distribution. *Biometrics 9*, 354–86.
3. Bliss, C. I. and Fisher, R. A. [1958]. Fitting of negative binomial distribution to biological data. *Biometrics 9*, 176–200.
4. Cramer, H. [1946]. *Mathematical Methods of Statistics*, Princeton Univ. Press, Princeton, N. J.
5. Evans, D. A. [1953]. Experimental evidence concerning contagious distributions in ecology. *Biometrika 40*, 186–210.
6. Feller, William [1957]. *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, New York, N. Y.
7. Gurland, J. [1957]. Some interrelations among compound and generalized distributions. *Biometrika 44*, 265-68.
8. McGuire, J. U., Brindley, T. A. and Bancroft, T. A. [1957]. The distribution of European corn-borer larvae *Pyrausta Nubilalis* (HBN), in field corn. *Biometrics 13*, 65–78.
9. Neyman, J. [1939]. On a new class of "contagious" distributions applicable in entomology and bacteriology. *Ann. Math. Stat. 10*, 35–57.
10. Shenton, L. R. [1949]. On the efficiency of the method of moments and Neyman's Type A distribution. *Biometrika 36*, 450–54.

APPENDIX

A. Formulae to compute factorial cumulants using moments about the origin:

We first obtain formulae to compute the first four factorial moments $\mu_{[i]}$, $i = 1, \cdots, 4$ using moments about the origin μ_i and then obtain formulae to evaluate the first four factorial cumulants $\kappa_{[i]}$ $i = 1, \cdots, 4$ using these factorial moments.

As for the first objective we note that $\mu_{[i]} = E\{x(x-1)\cdots(x-i+1)\}$, $i = 1, 2, \cdots$. By expanding the product within the expectation sign and using the elementary properties of the expectation operator, we get

$$\mu_{[1]} = \mu_1', \qquad \mu_{[2]} = \mu_2' - \mu_1', \qquad \mu_{[3]} = \mu_3' - 3\mu_2' + 2\mu_1',$$

and

$$\mu_{[4]} = \mu_4' - 6\mu_3' + 11\mu_2' - 6\mu_1'. \tag{16}$$

As for the latter, we observe that if $u(t)$ and $\psi(t)$ denote the factorial moment generating function and the factorial cumulant generating function, then $\psi(t) = \log u(t)$. On differentiating the equation successively with respect to t at $t = 0$ and noting that $\kappa_{[i]} = \psi^{(i)}(0)$, we have

$$\kappa_{[1]} = \mu_{[1]}, \qquad \kappa_{[2]} = \mu_{[2]} - \mu_{[1]}^2,$$
$$\kappa_{[3]} = \mu_{[3]} - 3\mu_{[1]}\mu_{[2]} + 2\mu_{[1]}^3,$$

and

$$\kappa_{[4]} = \mu_{[4]} - 3\mu_{[2]}^2 - 4\mu_{[1]}\mu_{[3]} + 12\mu_{[1]}^2\mu_{[2]} - 6\mu_{[1]}^4. \tag{17}$$

B. To prove formula (4):

Since τ_1, τ_2, τ_3 are functions of λ_1, λ_2 and λ_3, let us write them more explicitly as $\tau_1(\lambda_1, \lambda_2, \lambda_3)$, $\tau_2(\lambda_1, \lambda_2, \lambda_3)$ and $\tau_3(\lambda_1, \lambda_2, \lambda_3)$ respectively. If $\hat\lambda_1, \hat\lambda_2, \hat\lambda_3$ are the estimates of $\lambda_1, \lambda_2, \lambda_3$, then using the statistics t_1, t_2, t_3 which estimate the τ's consistently, we have

$$t_i = \tau_i(\hat\lambda_1, \hat\lambda_2, \hat\lambda_3), \qquad i = 1, 2, 3.$$

Hence, we have the asymptotic relations

$$t_i - \tau_i = \left(\frac{\partial\tau_i}{\partial\lambda_1}, \frac{\partial\tau_i}{\partial\lambda_2}, \frac{\partial\tau_i}{\partial\lambda_3}\right)(\hat\lambda_1 - \lambda_1, \hat\lambda_2 - \lambda_2, \hat\lambda_3 - \lambda_3)', \quad i = 1, 2, 3.$$

which can be rewritten as

$$\begin{bmatrix} \hat\lambda_1 - \lambda_1 \\ \hat\lambda_2 - \lambda_2 \\ \hat\lambda_3 - \lambda_3 \end{bmatrix} = \left[\frac{\partial(\tau_1, \tau_2, \tau_3)}{\partial(\lambda_1, \lambda_2, \lambda_3)}\right]^{-1} \begin{bmatrix} t_1 - \tau_1 \\ t_2 - \tau_2 \\ t_3 - \tau_3 \end{bmatrix}.$$

The generalized variance G of $\hat\lambda_1, \hat\lambda_2, \hat\lambda_3$ is then given by

$$\begin{aligned} G &= | E(\lambda_1 - \lambda_1, \lambda_2 - \lambda_2, \lambda_3 - \lambda_3)'(\hat\lambda_1 - \lambda_1, \hat\lambda_2 - \lambda_2, \hat\lambda_3 - \lambda_3) | \\ &= | V(t_1, t_2, t_3) | \Big/ \left| \frac{\partial(\tau_1, \tau_2, \tau_3)}{\partial(\lambda_1, \lambda_2, \lambda_3)} \right|^2 \end{aligned}$$

which is formula (4).

C. Equations for estimation:

We give below the equations for estimating the parameters for the various methods mentioned in Section 4. Their derivations are omitted for brevity.

i. Equations for the estimation of parameters using the first three moments are:

$$k = \frac{\hat{\kappa}_{[3]}\hat{\kappa}_{[1]}}{\hat{\kappa}_{[3]}\hat{\kappa}_{[1]} - (\hat{\kappa}_{[2]})^2} - 2, \tag{18}$$

$$p = \hat{\kappa}_{[2]}/(\hat{\kappa}_{[1]}(k + 1)), \tag{19}$$

$$\lambda = \hat{\kappa}_{[1]}/(kp). \tag{20}$$

ii. Equations for the estimation of parameters using the first two moments and the proportion $\hat{P}_0$ of zeros are:

$$p \log \left\{1 + \left(\frac{\hat{\kappa}_{[2]}}{\hat{\kappa}_{[1]}} - p\right) \frac{\log \hat{P}_0}{\hat{\kappa}_{[1]}}\right\} + \left(\frac{\hat{\kappa}_{[2]}}{\hat{\kappa}_{[1]}} - p\right) \log (1 + p), \tag{21}$$

$$k = \frac{\hat{\kappa}_{[2]}}{p\hat{\kappa}_{[1]}} - 1, \qquad \lambda = \hat{\kappa}_{[1]}/(kp). \tag{22}$$

iii. Equation for the estimation of the parameter p using the first two moments and the ratio $\hat{P}_1/\hat{P}_0$ of the first two frequencies is

$$\frac{\log (1 + p)}{p} = \frac{\hat{\kappa}_{[1]}}{\hat{\kappa}_{[2]}} \log \{\hat{\kappa}_{[1]}\hat{P}_0/\hat{P}_1\}. \tag{24}$$

k and λ are then estimated by using equations (22) and (23).

SOME RANK SUM MULTIPLE COMPARISONS TESTS

ROBERT G. D. STEEL
Institute of Statistics, North Carolina State College
Raleigh, North Carolina, U. S. A.

1. SUMMARY

Rank sum tests for multiple comparisons and suitable to the completely random design with equal sample sizes are discussed. Significance tables and some facts about the joint distribution of rank sums are given. An example illustrating test procedures and making use of the tables is presented.

2. INTRODUCTION

A number of nonparametric tests based on ranks have been proposed for the comparison of treatments in a completely random design. For example, we have the Wilcoxon-Mann-Whitney test [21, 10], basically a two-sample test with a per-comparison error rate.

Also, Kruskal and Wallis [9] have proposed a rank test which is an analogue of Snedecor's F-test. This test provides evidence concerning the presence of real differences but is of limited use in locating them.

Steel [16, 17] has presented rank tests for comparing treatments against control and for all pairwise comparisons. Both of these tests use experiment-wise error rates.

Pfanzagl [13], as part of a more general theory, has discussed a two-step nonparametric decision process based on ranks, for testing the null hypothesis that k samples come from the same population and, if this is rejected, for deciding which one of the samples comes from a different population. No tables are given but it is suggested that they might be obtained by random sampling. It is also shown that the limiting distribution of the multivariate criterion is multinormal.

The per-comparison error rate test is sometimes criticized, particularly when all possible paired comparisons are made, because it will almost certainly lead to false declarations of significance when the experiment includes many treatments and if customary significance levels are used. It is also deemed inappropriate when the experiment is considered to be the conceptual unit.

The experimentwise error rate test is sometimes criticized because it requires such a large difference to be declared significant that it becomes difficult to detect any but the largest real differences when customary significance levels are used. Also, it may be that the individual comparison is considered to be the conceptual unit.

A brief discussion of these error rates is given by Steel [18] in response to *Biometrics* Query 163.

Choice of a definition of error rate in the conduct of a particular experiment seems somewhat less crucial when it is realized that we can compute the significance level for a particular definition of error rate from knowledge of the chosen significance level for any other definition of error rate. This is not generally a simple computation unless the comparisons are independent. In the case of p independent comparisons, if α' and α are the experimentwise and per-comparison error rates respectively, we have the relation: $1 - \alpha' = (1 - \alpha)^p$.

When comparisons are not independent, computation of comparable significance levels for different definitions of error rate depends upon the extent of the dependence and the nature of the multiple comparisons test. No individual is likely to perform such a computation for a single experiment. Thus a table needs to be prepared for comparing the customary significance levels for differently defined error rates. For tests based on an underlying normal distribution, this has been done fairly extensively by Harter [4, 6].

The experimenter may try to meet the usual criticisms of per-comparison and experimentwise error rates by choice of a non-standard significance level or an alternative test procedure. Presently, tables of significant values for such levels do not appear to be available for experimentwise error rates; in the case of alternative tests, several are available, including the Newman-Keuls [11, 7] procedure and Duncan's [1,5] new multiple range test, which are sequential in nature.

This paper is concerned with rank tests, in particular with tables for an all-pairs-of-treatments test with an experimentwise error rate, and the use of these tables for a fixed rank sum test, an analogue of Tukey's w-procedure (for example, see Steel and Torrie [19]), and for two multiple rank sum tests, analogues of the Newman-Keuls procedure and of Duncan's test. Table 2 is used in the first two cases, Table 3 in the last case.

3. CONSTRUCTION OF TABLES

The proposed tests call for rank sums and their conjugates computed for all pairwise comparisons of treatments. The minimum of each rank sum and conjugate is used, the set of minima providing a multivariate rank sum test criterion. These sums are compared with a single tabulated value for the fixed rank sum test and with several values for the multiple rank sum tests. Table 2 provides critical values for the analogues of Tukey's and the Newman-Keuls tests; Table 3 provides for the analogue of Duncan's test.

Methods for constructing probability tables and limited tables have been presented earlier [16, 17]. Construction of exact tables of any extent was beyond the computing facilities available. However, some machine time was available and this was used for some sampling experiments.

It was originally intended to ignore the discrete nature of the data and to use the Kolmogorov-Smirnov [8, 14] one-sample test to determine the sample size necessary to attain a certain precision in the constructed tables. However, available computing facilities limited sampling to values of $k = 3$ and 4 (2 and 3 treatments when one was control) and $n = 4$ (1) 10. In addition, samples were obtained for $k = 5$, $n = 4$, 5 and $k = 6$, $n = 5$. The number of permutations obtained for each case was either 5000 or 6000. These tables were used only for checking purposes against the few exact distributions available, $k = 3$ and $n = 3$ (1) 6, and against approximations used in constructing these and earlier tables.

It was assumed that the various multivariate rank sum criteria are distributed approximately as multinormal distributions having mean vectors and variance-covariance matrices as given in the appendix. (Fraser's [3] vector form of the Wald-Wolfowitz-Hoeffding theorem does not apply since the $\| C_{n\alpha}(i, j) \|$'s of Fraser do not exist for the test criteria used here.)

On this assumption, one naturally proceeds to base computation on presently available tables. Tables for Tukey's and Duncan's tests are the obvious choice for all-pairs tests. These tests are based on a multinormal distribution with $\rho^2 = n_i n_j/(n_h + n_i)(n_h + n_j)$ where the present distribution calls for $\rho^2 = n_i n_j/(n_h + n_i + 1)(n_h + n_j + 1)$, a small difference. The appropriate tables are, then, tables of the Studentized range with known variance, that is, infinite degrees of freedom. Table 22 of Pearson and Hartley [12] is such a table, gives percentage points of .10, .05 and .01, is appropriate for the first two tests, and was used in computing Table 2. Corrected tables for Duncan's test have been computed by Harter [5] and this table was used in computing Table 3, also for percentage points of .10, .05 and .01.

Table 2 was constructed by taking the integral part of $\mu - t\sigma/\sqrt{2}$, unless the decimal fraction was $> .9$, in which case the next higher integer was tabulated, where t was obtained from the distribution of w/σ, $w =$ range, Table 22 of Pearson and Hartley [12]. Since rank sums are essentially differences, it is necessary to introduce $\sqrt{2}$ into the denominator as shown. Tabulated rank sum values for $\alpha = .10$ differed in only two cases from values obtained by sampling. In particular, for $k = 4$, $n = 4$, no value is significant by sampling; for $k = 3$, $n = 6$, a rank sum of 25 is significant by sampling whereas 26 is not. Values for $\alpha = .05$ differed in no case. Values for $\alpha = .01$ ran lower than those obtained by sampling, the difference increasing with n to a value of two in three cases. Hence, it is reasonable to conclude that tabulated values of the rank sum are conservative (low) for $\alpha = .01$.

The first attempt to construct Table 3 led to values which tended to run high for $\alpha = .10$, correct for $\alpha = .05$, and low for $\alpha = .01$, relative to values found by sampling. For this reason, tabulated values are of $\mu - t\sigma/\sqrt{2}$ decreased by unity for $\alpha = .10$, as computed for $\alpha = .05$, and increased by unity for $\alpha = .01$. On this basis, tabulated values for $\alpha = .10$ appear to be low, hence conservative, when not in agreement with sampling results; in particular, 7 out of 19 tabulated values are one unit low. For $\alpha = .05$, 3 out of 19 values are one unit high. For $\alpha = .01$, 5 out of 19 values are one unit low, with three of these being for $n = 5$.

Tables have already been constructed [16], using Dunnett's [2] tables, for the treatments against control test. These tables agree well with the sampling results. In no case is there a difference of more than one in the value of the test criterion, with the tables most often giving the conservative (lower) value.

TABLE 1

FINAL WEIGHTS OF CHICKS AT SIX WEEKS (GRAMS) FOR VARIOUS SOURCES OF PROTEIN SUPPLEMENT

H Horse-bean	*L* Linseed Oil Meal	*Sb* Soybean Oil Meal	*Sf* Sunflower Seed Oil Meal	*M* Meat Meal	*C* Casein
179	309	243	423	325	368
160	229	230	340	257	390
136	181	248	392	303	379
227	141	327	339	315	260
217	260	329	341	380	404
168	203	250	226	153	318
108	148	193	320	263	352
124	169	271	295	242	359
143	213	316	334	206	216
140	257	267	322	344	222

4. USE OF TABLES

To illustrate the use of the tables, the data in Query 60 of *Biometrics* (15) are used. These are presented in Table 1. Since the test is presently unavailable for unequal sample sizes, only the first ten items in each treatment are used.

The following set of minimum rank sums is obtained by pairwise rankings, a minimum being the lesser of the rank sum T, and its conjugate, $T' = (2n + 1)n - T$; minimum treatment is the treatment for which the rank sum is minimum. Ties were assigned their average rank. This gives a multivariate criterion with 15 entries.

Comparison	*H, Sf*	*H, Sb*	*H, C*	*L, Sf*	*H, M*	*L, C*
Minimum Rank Sum	56	57	58	60	62	$64\frac{1}{2}$
Minimum Treatment	H	H	H	L	H	L

Comparison	*Sb, Sf*	*H, L*	*L, Sb*	*L, M*	*Sb, C*	*Sf, M*	*M, C*	*Sb, M*	*Sf, C*
Minimum Rank Sum	71	75	75	$75\frac{1}{2}$	80	82	84	100	103
Minimum Treatment	*Sb*	*H*	*L*	*L*	*Sb*	*Sf*	*M*	*Sb*	*Sf*

From Table 2, a rank sum of 67 is significant at the 5 percent level, $k = 6$, $n = 10$; hence six comparisons are declared significant.

The device, used with multiple comparisons procedures, of underlining treatments which cannot be distinguished by their means may be adapted to apply to rank sum procedures. Thus, from the test, it appears that H and L, and L, Sb, M, Sf and C form two groups as a first step; since L can be distinguished from Sf and C, the latter group becomes two, namely L, Sb and M, and Sb, M, Sf and C. We have:

$$\underline{H \quad L} \quad Sb \quad M \quad Sf \quad C$$

Ordering of Sb and M, and of Sf and C was done on the basis of rank sums for these paired comparisons though this does not imply that this is the only, or even the best, method.

This procedure is an analogue of Tukey's w-procedure [19]. The significance level is for an experimentwise error rate. Use of Wilcoxon's [21] two-sample test, with its per-comparison error rate, calls for a significant rank sum of 78; this will result in four more comparisons being declared significant.

TABLE 2
Percentage Points of the Minimum Rank Sum
(an approximation)

Number in treatment	α	k = number of treatments being tested							
		3	4	5	6	7	8	9	10
4	.10	10	10	—	—	—	—	—	—
	.05	—	—	—	—	—	—	—	—
	.01	—	—	—	—	—	—	—	—
5	.10	17	16	15	15	—	—	—	—
	.05	16	15	—	—	—	—	—	—
	.01	—	—	—	—	—	—	—	—
6	.10	26	24	23	22	22	21	21	—
	.05	24	23	22	21	—	—	—	—
	.01	—	—	—	—	—	—	—	—
7	.10	36	34	33	32	31	30	30	29
	.05	34	32	31	30	29	28	28	—
	.01	29	28	—	—	—	—	—	—
8	.10	48	46	44	43	42	41	40	40
	.05	45	43	42	40	40	39	38	38
	.01	40	38	37	36	—	—	—	—
9	.10	62	59	57	56	55	54	53	52
	.05	59	56	54	53	52	51	50	49
	.01	52	50	48	47	46	45	—	—
10	.10	77	74	72	70	69	68	67	66
	.05	74	71	68	67	66	64	63	63
	.01	66	63	62	60	59	58	57	56

It is also possible to propose and carry out a sequential procedure, an analogue of the Newman-Keuls procedure, which uses several rank sums for testing. For this procedure, the above analysis is the first step and has separated the treatments into three groups. To proceed, we assume that declared differences are indeed real. Hence to test H versus L, the first group, we may use Wilcoxon's [21] two-sample test, H and L are declared significantly different and the line beneath them may be removed.

Further, compare L, Sb and M using the critical value for $k = 3$, $n = 10$, namely 74. L versus Sb and L versus M at 75 and $75\frac{1}{2}$ are beyond the 10 percent point but are not quite significant. We cannot distinguish among these three treatments.

Finally, consider the group composed of treatments Sb, M, Sf and C. The critical value is 71, $k = 4$, $n = 10$. The treatments Sb and Sf can be distinguished and we must, then, change the order of Sf and C from that proposed when the fixed rank sum test was used. No further differences will be declared significant by this procedure. We have:

$$H \qquad L \qquad Sb \qquad M \qquad C \qquad Sf$$

For the Tukey and Newman-Keuls parametric procedures, we find means of 160.2(H), 211.0(L), 267.4(Sb), 278.8(M), 323.2(Sf) and 326.8(C). Also $s_{\bar{x}} = 18.03$ and significant ranges are 51.2, 61.5, 67.6, 71.9 and 75.2 for $k = 2, \cdots, 6$ respectively.

For Tukey's test, the fixed rank sum is 75.2. We find:

$$H \qquad L \qquad Sb \qquad M \qquad Sf \qquad C$$

For the Newman-Keuls procedure, we find:

$$H \qquad L \qquad Sb \qquad M \qquad Sf \qquad C$$

We now compare the results obtained from applying the parametric and non-parametric procedures.

The fixed rank sum test and Tukey's test lead to the same conclusions. Both are based on experimentwise error rates.

Conclusions drawn from the multiple rank sum test and the Newman-Keuls test differ as follows. L versus M is significant by the Newman-Keuls procedure only; Sb versus Sf is significant by the rank sum test only. Otherwise, the procedures lead to the same conclusions. Fifteen paired tests have been made. Since L versus M is nearly significant and Sb versus Sf is just significant by the multiple rank sum test, it would appear that the two methods lead to conlcusions, for this example, that differ only slightly.

The other multiple rank sum test to be considered is an analogue of Duncan's new multiple range test. Table 3 provides critical values. A rank sum of 75 is significant at the 5 percent level, $k = 6$, $n = 10$; hence nine comparisons are significant. Tentatively, we have:

$$H \qquad L \qquad M \qquad Sb \qquad C \qquad Sf$$

Unfortunately, this includes an anomaly since Sb versus Sf is also declared significant. The same would be true if the Wilcoxon-Mann-Whitney test were being used at a significance level (between 5 percent and 1 percent) calling for a critical value of 75.

TABLE 3

Percentage Points of the Minimum Rank Sum for Duncan Analogue (an approximation)

Number in treatment	α	k = number of treatments being tested							
		3	4	5	6	7	8	9	10
4	.10	11	10	10	10	10	10	10	10
	.05	10	10	10	10	10	10	10	10
	.01	—	—	—	—	—	—	—	—
5	.10	18	18	17	17	17	17	17	17
	.05	17	17	16	16	16	16	16	16
	.01	15	15	15	—	—	—	—	—
6	.10	27	26	26	26	26	25	25	25
	.05	26	25	25	25	25	24	24	24
	.01	23	22	22	22	22	21	21	21
7	.10	37	37	37	36	36	36	36	36
	.05	36	35	35	35	34	34	34	34
	.01	32	32	31	31	30	30	30	30
8	.10	50	49	49	49	48	48	48	48
	.05	48	47	47	46	46	46	46	45
	.01	43	42	42	41	41	41	41	40
9	.10	65	64	63	63	62	62	62	62
	.05	62	61	60	60	60	59	59	59
	.01	56	55	54	54	53	53	53	52
10	.10	81	80	79	79	78	78	78	77
	.05	77	76	76	75	75	74	74	74
	.01	70	69	68	68	67	67	67	66

On the basis that treatments declared significantly different are indeed so, we proceed to test treatments M, Sb, C and Sf using $k = 4$ and L and M using $k = 2$. The final result is:

H L Sb M C Sf

This result is the same as that obtained using the Wilcoxon-Mann-Whitney test.

Using Duncan's (parametric) new multiple range test, we find:

H L Sb M Sf C

Conclusions from Duncan's test and its rank sum analogue differ only in that the parametric test finds *Sb* versus *C* to be significant.

The multiple rank sum test can be adapted to apply to testing treatments versus control as well.

5. APPENDIX—THEORY

The problem is concerned with pairwise testing of treatments in the one-way classification or completely random design.

The test criteria are rank sums, computed as for Wilcoxon's [21] two sample test, for appropriate pairs of treatments. Rank sums will be referred to Tables 2 or 3 for testing rather than to White's [20] table for the Wilcoxon-Mann-Whitney test.

Two tests will be considered:

1. The all pairs test, in detail.
2. Treatments against control, rather briefly.

For the all pairs test, let X_i, $i = 1, \cdots, k$ be random variables measuring some characteristic for each of k samples or treatments. Let there be n_i observations on the i-th treatment. Computation of the test criterion requires us to:

1. Rank the X_i's and the X_j's, all $i < j$, assigning rank 1 to the least observation.
2. Add ranks for the variable with fewer observations to give T_{ij}. (There is no loss of generality if we assume $n_1 \leq n_2 \leq \cdots \leq n_k$).
3. Compute the conjugate of T_{ij}, namely $T'_{ij} = (n_i + n_j + 1)n_i - T_{ij}$.

The conjugate is the rank total that would be obtained if rank 1 were assigned to the highest observation. Conjugates are required for two-tailed tests.

Consider $(T_{12}, \cdots, T_{1k}, T_{23}, \cdots, T_{k-1,k})$, a criterion with $\binom{k}{2}$ components. Rank tests are based on the assumption that, under H_0, all permutations of the $\sum n_i$ observations are equally likely. Hence, we must know the number of ways in which $(T_{12}, \cdots, T_{k-1,k})$ can be obtained. For this, a recursion formula is given in [14]. This provides a method, though tedious, of deriving the distribution of $(T_{12}, \cdots, T_{k-1,k})$.

The distribution of $(T_{12}, \cdots, T_{k-1,k})$ has been shown to have the following parameters [14]:

$$E(T_{ij}) = \mu_{ij} = n_i(n_i + n_j + 1)/2,$$

$$E(T_{ij} - \mu_{ij})^2 = \sigma^2_{ij} = n_i n_j(n_i + n_j + 1)/12,$$

$$E(T_{hi} - \mu_{hi})(T_{hj} - \mu_{hj}) = \sigma_{hi,hj} = n_h n_i n_j/12 = \sigma_{hj,ij},$$

$$E(T_{hi} - \mu_{hi})(T_{ij} - \mu_{ij}) = \sigma_{hi,ij} = -n_h n_i n_j/12,$$

$$E(T_{gh} - \mu_{gh})(T_{ij} - \mu_{ij}) = \sigma_{gh,ij} = 0,$$

$$\rho^2_{ih,hj} = \rho^2_{ih,jh} = \rho^2_{hi,hj} = n_i n_j/(n_h + n_i + 1)(n_h + n_j + 1),$$

$$\rho^2_{gh,ij} = 0.$$

The determinant of the variance-covariance matrix is:

$$[\prod_i n_i^{k-1}(\sum n_i + 1)^{k-1}]/12^{\binom{k}{2}}.$$

The elements in the inverse of the variance-covariance matrix are: Corresponding to the variance of T_{ij} ,

$$12(\sum n_\alpha + 1 - n_i - n_j)/n_i n_j(\sum n_\alpha + 1).$$

Corresponding to a covariance with the T's having a common subscript, h, in the same position:

$$-12/n_h(\sum n_\alpha + 1).$$

Corresponding to a covariance with the T's having a common subscript, i, in different positions, for example T_{hi} and T_{ij} :

$$12/n_i(\sum n_\alpha + 1).$$

Finally, the element corresponding to $\sigma_{gh,ij}$ is zero.

The determinant of the variance-covariance matrix may be evaluated as follows: from the ith row of the determinant, factor $n_1 n_{1+i}/12$, $i = 1, \cdots, k - 1$; from the $([k - 1] + i)$-th row, factor $n_2 n_{2+i}/12$, $i = 1, \cdots, k - 2$; $\cdots$; from the last row, factor $n_{k-1}n_k/12$. The product of these factors is $\prod_i n_i^{k-1}/12^{\binom{k}{2}}$.

The entries in the determinant which is the other factor may be described somewhat crudely as follows.

The i-th diagonal block, $i = 1, \cdots, k - 1$, which contains the variances and covariances of the T_{ij}'s, fixed i and $j > i$, will be

$$\begin{bmatrix} n_i + n_{i+1} + 1 & n_{i+2} & \cdots & n_k \\ n_{i+1} & n_i + n_{i+2} + 1 & \cdots & n_k \\ \cdots & \cdots & \cdots & \cdots \\ n_{i+1} & n_{i+2} & \cdots & n_i + n_k + 1 \end{bmatrix}.$$

This is a $(k - i) \times (k - i)$ block.

The block consisting of the same rows and the first $(k - 1)$ columns is

$$\left[\begin{array}{ccccccccc} 0 & \cdots & 0 & -n_1 & n_1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & -n_1 & 0 & n_1 & 0 & \cdots & 0 \\ & \cdots & & \cdots & & & \cdots & & \\ 0 & \cdots & 0 & -n_1 & 0 & 0 & 0 & \cdots & n_1 \\ \multicolumn{3}{c}{\underbrace{\qquad\qquad\qquad}_{i - 2 \text{ columns}}} & & \multicolumn{5}{c}{\underbrace{\qquad\qquad\qquad\qquad\qquad}_{k - i \text{ columns}}} \end{array}\right].$$

The next block to the right consists of $k - 2$ columns and is:

$$\begin{bmatrix} 0 & \cdots & 0 & -n_2 & n_2 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & -n_2 & 0 & n_2 & 0 & \cdots & 0 \\ & \cdots & & \cdots & & & \cdots & & \\ 0 & \cdots & 0 & -n_2 & 0 & 0 & & \cdots & n_2 \\ \multicolumn{3}{c}{\underbrace{\hphantom{0 \cdots 0}}_{i-3\text{ columns}}} & & \multicolumn{5}{c}{\underbrace{\hphantom{0\ 0\ 0 \cdots n_2}}_{k-i\text{ columns}}} \end{bmatrix}.$$

The pattern is now clear.

The first block to the right of the i-th diagonal block consists of the next $(k - [i + 1])$ columns and is

$$\begin{bmatrix} -n_{i+2} & -n_{i+3} & \cdots & -n_k \\ n_{i+1} & 0 & \cdots & 0 \\ 0 & n_{i+1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & n_{i+1} \end{bmatrix}.$$

The block to the right of this is

$$\begin{bmatrix} 0 & 0 & \cdots & 0 \\ -n_{i+3} & -n_{i+4} & \cdots & -n_k \\ n_{i+2} & 0 & \cdots & 0 \\ 0 & n_{i+2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & n_{i+2} \end{bmatrix}.$$

Again, the pattern is clear.

At the next step in evaluating the determinant, from column 1 subtract columns $([k-1]+1)$ through $([k-1]+[k-2])$. To column 2, add column $([k - 1] + 1)$ and subtract columns $([k - 1] + [k - 2] + 1)$ through $([k - 1] + [k - 2] + [k - 3])$. In general, to column i, $i = 1, \cdots, k - 1$, add the $i - 1$ columns described as those containing the first column of the diagonal blocks previously set out, and subtract the sum of the next $(k - [i + 1])$ columns, that is, the columns including the $(i + 1)$-st diagonal block.

The result of the above operations is that the first diagonal block has $\sum n_i + 1$ as common diagonal element and zeros elsewhere. The other blocks containing the first $k - 1$ columns are of the form previously described but with $-n_1$ and n_1 replaced by $-(\sum n_i + 1)$ and $\sum n_i + 1$ respectively.

At the next and final step, leave rows $1, \cdots, k - 1$ unaltered.

Consider the next $k - 2$ rows which we now call the first set, the next $k - 3$ rows called the second set, and so on, the i-th set consisting of $(k - [i + 1])$ rows, beginning with row $\sum_{\alpha=1}^{i} (k - \alpha) + 1$. To the j-the row of the i-th set, add row i and subtract row $i + j$, these rows coming from the first $k - 1$ rows of the determinant.

The result of these operations is a determinant with $\sum n_i + 1$ in the first $k - 1$ principal diagonal positions, ones elsewhere in the principal diagonal, and zeros below the principal diagonal. Hence the determinant is as given.

That the inverse elements are correctly given may be checked by multiplying the matrix by the given inverse.

For the treatments versus control procedure, let X_i, $i = 0, 1, \cdots, k$ be random variables measuring some characteristic of a control and k treatments with n_i observations in the i-th sample.

Computation of the test criterion requires us to:

1. Rank jointly the X_0's and X_i's, i fixed, giving rank 1 to the least observation.
2. Add ranks for the variable with fewer observations, here assumed to be the check, to give T_i.
3. Compute the conjugate, T_i'.

Consider $(T_1, \cdots, T_k)$. Again, a recursion formula for finding the number of permutations which give rise to a specific value of $(T_1, \cdots, T_k)$ is given in [13].

The distribution of $(T_1, \cdots, T_k)$ has been shown to have the following parameters [13]:

$$E(T_i) = \mu_i = n_0(n_0 + n_i + 1)/2,$$
$$E(T_i - \mu_i)^2 = \sigma_i^2 = n_0 n_i (n_0 + n_i + 1)/12,$$
$$E(T_i - \mu_i)(T_j - \mu_j) = \sigma_{ij} = n_0 n_i n_j/12,$$
$$\rho_{ij}^2 = n_i n_j/(n_0 + n_i + 1)(n_0 + n_j + 1).$$

It may also be shown that the determinant of the variance-covariance matrix is

$$\prod_0^k n_i (n_0[n_0 + 1])^{k-1} \left(\sum_0^k n_i + 1 \right) / 12.$$

The diagonal and off-diagonal elements of the inverse are, respectively,

$$12 \left(\sum_{\alpha \neq i} n_\alpha + 1 \right) \Big/ \left[n_0(n_0 + 1) n_i \left(\sum_0^k n_\alpha + 1 \right) \right],$$

and

$$-12 \Big/ \left[n_0(n_0 + 1) \left(\sum_0^k n_i + 1 \right) \right].$$

To evaluate the determinant of the variance-covariance matrix, factor $n_0 n_i/12$ from the i-th row of the determinant. This gives:

$$\frac{n_0^{k-1} \prod_0^k n_i}{12^k} \begin{vmatrix} n_0 + n_1 + 1 & n_2 & \cdots & n_k \\ n_1 & n_0 + n_2 + 1 & \cdots & n_k \\ \cdots & \cdots & \cdots & \cdots \\ n_1 & n_2 & \cdots & n_0 + n_k + 1 \end{vmatrix}.$$

Next, subtract the k-th row from the i-th row, $i = 1, \cdots, k - 1$. We obtain

$$\frac{n_0^{k-1} \prod_0^k n_i}{12^k} \begin{vmatrix} n_0 + 1 & 0 & \cdots & -(n_0 + 1) \\ 0 & n_0 + 1 & \cdots & -(n_0 + 1) \\ \cdots & \cdots & \cdots & \cdots \\ n_1 & n_2 & \cdots & n_0 + n_k + 1 \end{vmatrix}.$$

Finally, obtain a new k-th column as the sum of all columns. It is then apparent that the given determinant is correct.

That the given inverse elements are correct is seen by multiplying the matrix by the stated inverse.

From the above information, it is possible to tabulate probabilities for the parent distribution and derived distributions of interest.

6. ACKNOWLEDGEMENT

I wish to thank Dr. Richard C. Lesser, Director, Cornell Computing Center, Cornell University, Ithaca, New York, for making machine time and programming help available for the random sampling.

A considerable portion of the work in this paper was done while I was a member of the Biometrics Unit, Department of Plant Breeding, Cornell University.

7. REFERENCES

[1] Duncan, D. B. [1955]. Multiple range and multiple F tests. *Biometrics 11,* 1–42.

[2] Dunnett, C. W. [1955]. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Stat. Assoc. 50,* 1096–1121.

[3] Fraser, D. A. S. [1956]. A vector form of the Wald-Wolfowitz-Hoeffding theorem. *Ann. Math. Stat. 27,* 540–43.

[4] Harter, H. L. [1957]. Error rates and sample sizes for range tests in multiple comparisons. *Biometrics 13,* 511–36.

[5] Harter, H. L. [1960]. Critical values for Duncan's new multiple range test. *Biometrics 16,* 671–85.

[6] Harter, H. L. [1961]. Note 161-Corrected error rates for Duncan's new multiple range test. *Biometrics 17,* 321–24.

[7] Keuls, M. [1952]. The use of the 'studentized range' in connection with an analy sis of variance. *Euphytica 1,* 112–22.

[8] Kolmogorov, A. [1941]. Confidence limits for an unknown distribution function. *Ann. Math. Stat. 12*, 461–63.

[9] Kruskal, W. H., and W. A. Wallis. [1952]. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc. 47*, 583–621.

[10] Mann, H. B., and D. R. Whitney. [1947]. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat. 18*, 50–60.

[11] Newman, D. [1939]. The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika 31*, 20–30.

[12] Pearson, E. S., and H. O. Hartley. [1954]. *Biometrika Tables for Statisticians*, Vol. I. Cambridge University Press.

[13] Pfanzagl, J. [1959]. Ein kombiniertes Test und Klassifikations—Problem. *Metrika 2*, 11–45.

[14] Smirnov, N. [1948]. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat. 19*, 279–81.

[15] Snedecor, G. W. [1948]. Answer to Query 60. *Biometrics 4*, 213–15.

[16] Steel, R. G. D. [1959]. A multiple comparison rank sum test: Treatments versus control. *Biometrics 15*, 560–72.

[17] Steel, R. G. D. [1960]. A rank sum test for comparing all pairs of treatments. *Technometrics 2*, 197–207.

[18] Steel, R. G. D. [1961]. Answer to Query 163. *Biometrics 17*, 326–28.

[19] Steel, R. G. D., and J. H. Torrie. [1960]. *Principles and procedures of statistics*, McGraw-Hill Book Company, Inc., New York.

[20] White, C. [1952]. The use of ranks in a test of significance for comparing two treatments. *Biometrics 8*, 33–41.

[21] Wilcoxon, F. [1945]. Individual comparisons by ranking methods. *Biometrics 1*, 80–3.

THE ESTIMATION OF REPEATABILITY AND HERITABILITY FROM RECORDS SUBJECT TO CULLING

R. N. CURNOW[1]
Agricultural Research Council Unit of Statistics,
University of Aberdeen, Aberdeen, Scotland[2]

1. INTRODUCTION

In any animal breeding selection programme, estimates of repeatability and heritability are needed to choose between the various selection schemes available and also to ensure that the highest possible genetic gains are obtained from the chosen scheme. Estimates of repeatability and heritability are generally subject to large sampling errors. Therefore, the most efficient methods of estimation should be used even if they do involve rather lengthy computations. In this paper, the maximum likelihood estimation of repeatability and heritability from records subject to culling will be considered. The more usual regression estimators are often very inefficient compared with these maximum likelihood estimators.

Suppose that we wish to estimate the repeatability of lactation yield in a herd of dairy cattle. We shall assume that only first and second lactation yields are available and that, if there had been no culling (*i.e.*, if all the cows had had second records as well as first records), the first and second records would have been normally distributed over the herd with means μ_1 and μ_2, variances σ_1^2 and σ_2^2 and covariance between the two records of the same cow σ_{12}. The first and second records of cow i will be written y_{i1} and y_{i2} respectively. The assumption of normality for the distributions will probably be a reasonably good approximation unless the herd can be split into groups so that any two cows in the same group are much more alike than two cows in different groups. These groups may, for example, be groups of daughters of the same sire or groups of cows according to the year in which they gave their first record or the month in which they calved. Methods are available for making allowances for such groupings, but they will be assumed absent in the rest of this paper. Very rarely will the culling

[1] The work on this paper was undertaken while the author was visiting the United States on a Harkness Fellowship of the Commonwealth Fund.

[2] Present address: Unit of Biometry, University of Reading.

in the herd be sufficiently intense or sufficiently highly correlated with future milk yielding capacity to affect seriously the normality of the distributions.

Repeatability is defined as the correlation between two different records of the same cow and is therefore

$$\rho = \sigma_{12}/\sigma_1\sigma_2 .$$

Since we are considering only first and second records, the question of whether the repeatability is the same for all pairs of records will not be discussed. The assumption is frequently made that $\sigma_1^2 = \sigma_2^2$ and, therefore, that repeatability is the same quantity as the regression coefficient of second records on first, $\beta_{21} = \sigma_{12}/\sigma_1^2$. The assumption is made, for example, in the formula given by Lush [1945] for comparing cows with a differing number of records and in the formula given by Lerner [1958] for the ratio of the heritability of the mean of n records to the heritability of a single record. It was also used by Henderson, Kempthorne, Searle and von Krosigk [1959] in their discussion of the disentanglement of environmental and genetic trends from records subject to culling. When $\sigma_1^2 \neq \sigma_2^2$, an estimate of β_{21} is needed for prediction purposes but for other purposes estimates of σ_1^2 and σ_2^2 may also be required. We shall assume in this paper that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, say, and therefore that $\beta_{21} = \rho$. A logarithmic transformation applied to the data may sometimes be useful in satisfying this assumption.

ρ is generally estimated by b, the sample regression coefficient of second records on first. There are two reasons for this. First, b is very simple to calculate and, second, it is an unbiassed estimator of ρ despite any culling that may have been based on the first records. However, when $\sigma_1^2 = \sigma_2^2$ and the first records of all cows are available, whether or not they have second records, b is not the maximum likelihood estimator of ρ. The variance of the second records about their regression on the first estimates $\sigma^2(1 - \rho^2)$ and the variance of all the first records estimates σ^2. These two estimators can be combined to give an estimator of ρ^2. Maximum likelihood makes use of this information as well as the information given by b.

We shall derive the maximum likelihood estimators of ρ and σ^2. We shall show that the efficiency of b as an estimator of ρ, relative to the maximum likelihood estimator, is fairly low for values of the various parameters that may well occur in practice. The bias of the maximum likelihood estimator is shown to be small. This suggests that the maximum likelihood estimator may be worth calculating. The computations involve only the solution of a cubic equation. A section is devoted to an approximate check of the assumption that $\sigma_1^2 = \sigma_2^2$.

Attention has been confined so far to the estimation of repeatability. The methods to be discussed could also be applied to the estimation of heritability from parent-offspring records. The assumption $\sigma_1^2 = \sigma_2^2$ means that, had there been no selection of parents, the variance of the parents and of the offspring would have been equal. In heritability studies, the maximum likelihood method makes use of information about animals that are not parents of animals which also have records. In milk yield studies, this would include information on dams that have only male calves.

2. THE MAXIMUM LIKELIHOOD ESTIMATION OF REPEATABILITY

All the cows have first lactation records but they do not all have second lactation records. We shall assume that the probability that a cow has a second record depends on its first record but not on any other character correlated with the second record. This rules out culling based on information about relatives or on characters such as percentage butter-fat. However, the effect of such culling on the estimates will often be very small and could probably be safely ignored.

Let N cows have a first record and $n \leq N$ cows have a second record. Numbering the cows with a second record $i = 1, 2, \cdots, n$ and the cows without a second record $i = n + 1, n + 2, \cdots, N$, the first records can be written

$$y_{i1} \qquad (i = 1, 2, \cdots, N)$$

and the second records

$$y_{i2} \qquad (i = 1, 2, \cdots, n).$$

The N first records are normally distributed with mean μ_1 and variance σ^2. Because the culling is based only on first records, the distribution of a second record y_{i2}, given the first y_{i1}, is independent of the distribution of the first record and is normal with mean $\mu_2 + \rho(y_{i1} - \mu_1)$ and variance $\sigma^2(1 - \rho^2)$. The likelihood of all the records is therefore

$$L = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{\sum_{i=1}^{N}(y_{i1} - \mu_1)^2}{2\sigma^2}\right\}$$
$$\times \frac{1}{[2\pi\sigma^2(1 - \rho^2)]^{n/2}} \exp\left\{-\frac{\sum_{i=1}^{n}[y_{i2} - \mu_2 - \rho(y_{i1} - \mu_1)]^2}{2\sigma^2(1 - \rho^2)}\right\}$$

and the log likelihood is, apart from a constant,

$$\ln L = -\frac{(n+N)}{2}\ln\sigma^2 - \frac{n}{2}\ln(1-\rho^2) - \frac{(1-\rho^2)\sum_{i=1}^{N}(y_{i1}-\mu_1)^2 + \sum_{i=1}^{n}[y_{i2}-\mu_2-\rho(y_{i1}-\mu_1)]^2}{2\sigma^2(1-\rho^2)}. \tag{2.1}$$

Kempthorne and von Krosigk (Henderson, Kempthorne, Searle and von Krosigk [1959]) suggested the use of maximum likelihood to estimate repeatability from records subject to culling. They derived the log likelihood when the cows were classified into groups and when the number of records for each cow was perfectly general. They derived very lengthy equations from which the maximum likelihood estimates of the parameters could be determined. In this paper, we are studying in greater detail the special case when there is no grouping of the cows and when there are at most only two records per cow.

(2.1) can be written

$$\begin{aligned}\ln L = -\frac{(n+N)}{2}\ln\sigma^2 - \frac{n}{2}\ln(1-\rho^2) - \frac{Ns^2}{2\sigma^2}& \\ - \frac{n(\rho^2\Sigma_1^2 + \Sigma_2^2 - 2\rho\Sigma_{12})}{2\sigma^2(1-\rho^2)}& \\ - \frac{N(\bar{y}-\mu_1)^2}{2\sigma^2} - \frac{n[\bar{y}_2-\mu_2-\rho(\bar{y}_1-\mu_1)]^2}{2\sigma^2(1-\rho^2)}&\end{aligned} \tag{2.2}$$

where

$$s^2 = \frac{\sum_{i=1}^{N}(y_{i1}-\bar{y})^2}{N}, \qquad \Sigma_1^2 = \frac{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2}{n},$$

$$\Sigma_2^2 = \frac{\sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2}{n}, \qquad \Sigma_{12} = \frac{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2)}{n},$$

$$\bar{y} = \frac{\sum_{i=1}^{N}y_{i1}}{N}, \qquad \bar{y}_1 = \frac{\sum_{i=1}^{n}y_{i1}}{n} \quad\text{and}\quad \bar{y}_2 = \frac{\sum_{i=1}^{n}y_{i2}}{n}.$$

The maximum likelihood estimators of μ_1 and μ_2 are clearly given by

$$\hat{\mu}_1 = \bar{y} \quad\text{and}\quad \hat{\mu}_2 = \bar{y}_2 - \hat{\rho}(\bar{y}_1 - \mu_1).$$

Also

$$\begin{aligned}\frac{\partial(\ln L)}{\partial(\sigma^2)} = -\frac{(n+N)}{2\sigma^2} + \frac{Ns^2}{2\sigma^4} + \frac{n(\rho^2\Sigma_1^2+\Sigma_2^2-2\rho\Sigma_{12})}{2\sigma^4(1-\rho^2)}& \\ + \frac{N(\bar{y}-\mu_1)^2}{2\sigma^4} + \frac{n[\bar{y}_2-\mu_2-\rho(\bar{y}_1-\mu_1)]^2}{2\sigma^4(1-\rho^2)}&\end{aligned} \tag{2.3}$$

and

$$\frac{\partial(\ln L)}{\partial \rho} = \frac{n\rho}{1 - \rho^2} - \frac{n[\rho(\Sigma_1^2 + \Sigma_2^2) - (1 + \rho^2)\Sigma_{12}]}{\sigma^2(1 - \rho^2)^2} + \frac{n(\bar{y}_1 - \mu_1)[\bar{y}_2 - \mu_2 - \rho(\bar{y}_1 - \mu_1)]}{\sigma^2(1 - \rho^2)} - \frac{n\rho[\bar{y}_2 - \mu_2 - \rho(\bar{y}_1 - \mu_1)]^2}{\sigma^2(1 - \rho^2)^2}. \tag{2.4}$$

By setting (2.3) and (2.4) equal to zero and substituting the maximum likelihood estimators of μ_1 and μ_2, $\hat{\rho}$ satisfies the cubic

$$(Ns^2 - n\Sigma_1^2)\hat{\rho}^3 - (N - n)\Sigma_{12}\hat{\rho}^2 + [(n + N)\Sigma_1^2 - N(s^2 - \Sigma_2^2)]\hat{\rho} - (n + N)\Sigma_{12} = 0. \tag{2.5}$$

This equation will have one and only one root between -1 and $+1$ of the same sign as the simple estimator $b = \Sigma_{12}/\Sigma_1^2$. There may be two other roots between -1 and $+1$ of opposite sign to b. The formula for $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{\hat{\rho}(\Sigma_1^2 + \Sigma_2^2) - (1 + \hat{\rho}^2)\Sigma_{12}}{\hat{\rho}(1 - \hat{\rho}^2)}. \tag{2.6}$$

When there is no culling, $s^2 = \Sigma_1^2$ and

$$\hat{\rho} = \frac{2\Sigma_{12}}{\Sigma_1^2 + \Sigma_2^2}.$$

When $\sigma_1^2 = \sigma_2^2$, the maximum likelihood estimator of the correlation coefficient has the arithmetic rather than the geometric mean of the two sample variances in the denominator.

To obtain the asymptotic variance of $\hat{\rho}$, we need the expected values of the second-order partial derivatives of $\ln L$ with respect to μ_1, μ_2, ρ and σ^2. These can be derived from a knowledge of the expected values of s^2, Σ_2^2 and Σ_{12}. Σ_1^2 depends on the method of culling and so will be taken as fixed. For reasons to be given later, we shall calculate the expected values when $\sigma_1^2 \neq \sigma_2^2$. Clearly, $E(s^2) = (N - 1/N)\sigma_1^2$. By writing

$$y_{i2} = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(y_{i1} - \mu_1) + \sigma_2\sqrt{1 - \rho^2}\, e_{i2.1},$$

so that $e_{i2\ 1}$ has a standard normal distribution independent of y_{i1},

$$\Sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\rho\sigma_2}{\sigma_1}(y_{i1} - \bar{y}_1) + \sigma_2\sqrt{1 - \rho^2}\,(e_{i2.1} - \bar{e}_{.2.1})\right]^2,$$

and, therefore, for given Σ_1^2,

$$E(\Sigma_2^2) = \frac{\rho^2 \sigma_2^2}{\sigma_1^2} \Sigma_1^2 + \frac{(n-1)}{n} \sigma_2^2(1-\rho^2).$$

Similarly,

$$E(\Sigma_{12}) = \frac{\rho\sigma_2}{\sigma_1} \Sigma_1^2 .$$

Also, with $\sigma_1^2 = \sigma_2^2$, $E(\bar{y} - \mu_1)^2 = \sigma^2/N$, $E[\bar{y}_2 - \mu_2 - \rho(\bar{y}_1 - \mu_1)]^2 = \sigma^2(1-\rho^2)/n$ and $E\{(\bar{y}_1 - \mu_1)[\bar{y}_2 - \mu_2 - \rho(\bar{y}_1 - \mu_1)]\} = 0$. After considerable algebra the asymptotic variance—covariance matrix for $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}^2$ and $\hat{\rho}$ is found to be **V**, where

$$\mathbf{V}^{-1} = \begin{bmatrix} \dfrac{N}{\sigma^2} + \dfrac{n\rho^2}{\sigma^2(1-\rho^2)} & -\dfrac{n\rho}{\sigma^2(1-\rho^2)} & 0 & -\dfrac{n\rho(\bar{y}_1 - \mu_1)}{\sigma^2(1-\rho^2)} \\[2ex] -\dfrac{n\rho}{\sigma^2(1-\rho^2)} & \dfrac{n}{\sigma^2(1-\rho^2)} & 0 & \dfrac{n(\bar{y}_1 - \mu_1)}{\sigma^2(1-\rho^2)} \\[2ex] 0 & 0 & \dfrac{n+N}{2\sigma^4} & -\dfrac{n\rho}{\sigma^2(1-\rho^2)} \\[2ex] -\dfrac{n\rho(\bar{y}_1 - \mu_1)}{\sigma^2(1-\rho^2)} & \dfrac{n(\bar{y}_1 - \mu_1)}{\sigma^2(1-\rho^2)} & -\dfrac{n\rho}{\sigma^2(1-\rho^2)} & \theta \end{bmatrix}$$

and

$$\theta = \frac{n}{\sigma^2(1-\rho^2)^2}[(1-\rho^2)\Sigma_1^2 + 2\rho^2\sigma^2] + \frac{n(\bar{y}_1 - \mu_1)^2}{\sigma^2(1-\rho^2)}.$$

The asymptotic variance of $\hat{\rho}$ is

$$V(\hat{\rho}) = \frac{(N+n)(1-\rho^2)^2}{n[(N+n)(1-\rho^2)\Sigma_1^2/\sigma^2 + 2N\rho^2]}.$$

Writing $n/N = S$, so that $(1 - S)$ is the culling intensity, and $\Sigma_1^2/\sigma^2 = 1/c$, where c measures the effect of the culling on the variance of the first records and is therefore a possible measure of the efficiency of the culling,

$$V(\hat{\rho}) = \frac{1}{n} \frac{(1+S)(1-\rho^2)^2}{\left[\dfrac{(1+S)(1-\rho^2)}{c} + 2\rho^2\right]}. \tag{2.7}$$

We shall now derive the approximate bias in $\hat{\rho}$ as an estimator of ρ. By substituting $\hat{\rho} = \rho + \delta$ in (2.5) and ignoring terms in δ^2 and δ^3,

$$\delta \doteq -\frac{(Ns^2 - n\Sigma_1^2)\rho^3 - (N-n)\Sigma_{12}\rho^2 + [(n+N)\Sigma_1^2 - N(s^2 - \Sigma_2^2)]\rho - (n+N)\Sigma_{12}}{3(Ns^2 - n\Sigma_1{}^2)\rho^2 - 2(N-n)\Sigma_{12}\rho + [(n+N)\Sigma_1{}^2 - N(s^2 - \Sigma_2^2)]}. \tag{2.8}$$

Approximating $E(\delta)$ by taking expected values separately in the numerator and denominator,

$$E(\delta) \doteq -\frac{\rho(1-\rho^2)\sigma^2(n-N)}{\Sigma_1^2 n(n+N)(1-\rho^2)+\sigma^2[(n-N)+\rho^2(2nN+N-3n)]}.$$

To order $1/n$,

$$E(\delta) = \frac{\rho(1-\rho^2)(1-S)}{n\left[\dfrac{(1+S)(1-\rho^2)}{c}+2\rho^2\right]}. \tag{2.9}$$

Therefore, from (2.7), the approximate ratio of the bias to the standard deviation is

$$\frac{E(\delta)}{\sqrt{V(\hat{\rho})}} \doteq \frac{\rho(1-S)}{\sqrt{n}\,\sqrt{1+S}\left\{\dfrac{1+S}{c}(1-\rho^2)+2\rho^2\right\}^{1/2}}$$
$$\leq \frac{\rho(1-S)}{\sqrt{n}\,\sqrt{1+S}}\operatorname{Max}\left[\sqrt{\frac{c}{1+S}},\, 1/\sqrt{2}\right].$$

If n is reasonably large, the bias in $\hat{\rho}$ is unlikely to be serious. The bias will become relatively more important when estimates of ρ from different sources are pooled.

3. ESTIMATION OF σ_1^2 AND σ_2^2 WHEN $\sigma_1^2 \neq \sigma_2^2$

In the previous section we derived the expected values of s^2, Σ_2^2 and Σ_{12} for given Σ_1^2 when $\sigma_1^2 \neq \sigma_2^2$. They were

$$E(s^2) = \frac{N-1}{N}\sigma_1^2,$$

$$E(\Sigma_2^2) = \frac{\rho^2\sigma_2^2}{\sigma_1^2}\Sigma_1^2 + \frac{(n-1)}{n}\sigma_2^2(1-\rho^2),$$

and

$$E(\Sigma_{12}) = \frac{\rho\sigma_2}{\sigma_1}\Sigma_1^2.$$

Apart from the usual slight differences in the multipliers, important only for small n and N, the following estimators of σ_1^2, $\beta = \rho\sigma_2/\sigma_1$ and $\sigma_2^2(1-\rho^2)$ are the maximum likelihood estimators when $\sigma_1^2 \neq \sigma_2^2$,

$$\hat{\sigma}_1^2 = Ns^2/(N-1),$$

$$\hat{\beta}_{21} = b = \Sigma_{12}/\Sigma_1^2,$$

and

$$\widehat{\sigma_2^2(1 - \rho^2)} = \frac{n}{n - 2}\,[\Sigma_2^2 - \Sigma_{12}^2/\Sigma_1^2].$$

From the expected values of s^2, Σ_2^2 and Σ_{12} given in the previous section these estimators are all unbiassed. The maximum likelihood estimator of σ_2^2/σ_1^2 is

$$\widehat{\left(\frac{\sigma_2^2}{\sigma_1^2}\right)} = \frac{\Sigma_{12}^2}{\Sigma_1^4} + \frac{\Sigma_2^2 - \Sigma_{12}^2/\Sigma_1^2}{s^2}. \tag{3.1}$$

No exact method is available for constructing a confidence interval for σ_2^2/σ_1^2 except when $n = N$, *i.e.*, when there is no culling (see Curnow [1957] for references and for a method to be used when the data are grouped). The asymptotic variance of $\widehat{(\sigma_2^2/\sigma_1^2)}$ could be derived and used to provide an approximate confidence interval. This would give some indication of the importance of the assumption that $\sigma_1^2 = \sigma_2^2$. A simpler, but much less sensitive, test of whether $\sigma_2^2 > \sigma_1^2$ could be based on the fact that the quantity

$$\frac{(N - 1)n\sigma_1^2}{N(n - 2)\sigma_2^2(1 - \rho^2)} \times \frac{\Sigma_2^2 - \Sigma_{12}^2/\Sigma_1^2}{s^2},$$

which compares a χ^2-value from the variation of the second records about their regression on the first with a χ^2-value from the variation of all the first records, has an F-distribution with $n - 2$ and $N - 1$ degrees of freedom.

$$F = \frac{(N - 1)n}{N(n - 2)}\,\frac{\Sigma_2^2 - \Sigma_{12}^2/\Sigma_1^2}{s^2}$$

significantly greater than $F = 1$ would suggest that $\sigma_2^2(1 - \rho^2) > \sigma_1^2$ and, therefore, that $\sigma_2^2 > \sigma_1^2$.

4. THE EFFICIENCY OF THE SIMPLE REGRESSION ESTIMATOR OF REPEATABILITY

The statistic $b = \Sigma_{12}/\Sigma_1^2$ is very easy to calculate and is the one generally used to estimate repeatability. It is the only possible estimator if first records are available only for those cows also having second records. Providing that culling is based solely on the first lactation yields, b is always an unbiassed estimator of $\beta_{21} = \rho\sigma_2/\sigma_1$, but an unbiassed estimator of ρ only when $\sigma_1^2 = \sigma_2^2$. In this paper we are assuming $\sigma_1^2 = \sigma_2^2$. The variance of b is

$$V(b) = \sigma^2/n\Sigma_1^2 = c/n. \tag{4.1}$$

From (2.7), the asymptotic efficiency of b relative to the maximum likelihood estimator, $\hat{\rho}$, is therefore

$$Eff. = \frac{(1 + S)(1 - \rho^2)^2}{(1 + S)(1 - \rho^2) + 2c\rho^2}.$$

The values of this efficiency are shown in Table 1 for various values of S, ρ and c.

The following considerations suggest that c is unlikely to be greater than $c = 2$. Let the N cows be reduced to LN by accidental factors uncorrelated with the level of yield. Then a proportion $P = n/LN = S/L$ can be selected on the basis of first lactation yield. Assume that the selection is of the proportion S/L of the herd having the highest yields and that n and LN are sufficiently large that the effect of the selection

TABLE 1

THE ASYMPTOTIC EFFICIENCY OF THE SIMPLE REGRESSION ESTIMATOR OF REPEATABILITY RELATIVE TO THE MAXIMUM LIKELIHOOD ESTIMATOR

Overall Selection Intensity ($S = n/N$)	Repeatability (ρ)	$c = \sigma^2/\Sigma_1^2$				
		1/2	1	4/3	2	4
1/4	0	1.00	1.00	1.00	1.00	1.00
	1/4	0.95	0.90	0.88	0.82	0.70
	1/2	0.79	0.65	0.58	0.48	0.32
	3/4	0.49	0.33	0.27	0.20	0.11
	1	0	0	0	0	0
1/2	0	1.00	1.00	1.00	1.00	1.00
	1/4	0.96	0.92	0.89	0.85	0.74
	1/2	0.82	0.69	0.63	0.53	0.36
	3/4	0.54	0.37	0.30	0.23	0.13
	1	0	0	0	0	0
3/4	0	1.00	1.00	1.00	1.00	1.00
	1/4	0.96	0.93	0.91	0.87	0.77
	1/2	0.84	0.72	0.66	0.57	0.40
	3/4	0.58	0.40	0.34	0.25	0.15
	1	0	0	0	0	0
1	0	1.00	1.00	1.00	1.00	1.00
	1/4	0.97	0.94	0.92	0.88	0.79
	1/2	0.86	0.75	0.69	0.60	0.43
	3/4	0.61	0.44	0.37	0.28	0.16
	1	0	0	0	0	0

can be approximated by the truncation of the top P proportion of an infinite normal population with mean μ_1 and variance σ^2. The variance of the selected first records will then be

$$\Sigma_1^2 = \sigma^2[1 - \nu(\nu - T)],$$

where $\nu = Z/P$ and Z and T are the ordinate and abscissa at the point above which lies a proportion P of the population (Finney [1957] derives this formula and provides a table of values of $\nu' = \nu(\nu - T)$ for various values of P). The division of the overall selection intensity, S, between L and P is never complete and is certainly never known exactly. However, P is generally considerably larger than S, *i.e.*, there is a large amount of random culling and not so much actual selection. The value of c is

$$c = \frac{\sigma^2}{\Sigma_1^2} = 1/[1 - \nu(\nu - T)].$$

For various P, it takes the following values:

$P =$	0.95	0.9	0.8	0.7	0.6	0.5	0.1,
$c =$	1.24	1.40	1.72	2.03	2.37	2.75	5.91.

In practice, P is unlikely to be less than $P = 0.7$ and therefore c is unlikely to exceed $c = 2$. Table 1 does include $c = 4$ to illustrate the effect of intense selection based on yield and $c = \frac{1}{2}$ to illustrate the effect of a selection scheme that results in an increased rather than a decreased variance. The efficiency is the same for ρ as for $-\rho$ and so ρ is shown as positive in the table. In practice, ρ is very unlikely to be negative.

When $\rho = 0$, the efficiency of the simple regression estimator is 1. When $\rho = 1$, it is 0. For intermediate values of ρ, the efficiency increases rather slowly with S for fixed c, but decreases fairly rapidly with increasing c. For all values of c and S, the efficiency is very low for high values of ρ. As an example of the efficiency likely to occur in the estimation of the repeatability of lactation records, the efficiency is 0.66 when $c = \frac{4}{3}$, $\rho = \frac{1}{2}$ and $S = \frac{3}{4}$. In this case, the maximum likelihood estimator would certainly be worth calculating.

5. THE ESTIMATION OF REPEATABILITY BY AN ANALYSIS OF VARIANCE

Sometimes the effect of culling based on the first records is ignored and ρ estimated from a least squares analysis of variance of the $N + n$ first and second records. Table 2 shows this analysis of variance. The "cows and periods" sum of squares has been split into both "cows adjusting for periods and periods ignoring cows" and "cows ignoring periods and periods adjusting for cows". By a period difference is meant an average difference between first and second lactation records. The expected values of four important mean squares are shown. Δ_1, Δ_2 and Δ_3 are defined at the foot of the table and vanish only if the culling is random with respect to the first records. Two cases need to be distinguished. In the first we assume $\mu_1 = \mu_2$ and in the second we do not. We shall use the reference numbers at the right of the table. If $\mu_1 = \mu_2$, mean square 2 and a mean square obtained by pooling sums of squares 3 and 4 are used to estimate σ^2 and ρ. This is exactly equivalent to an analysis of a one way classification with unequal sub-class numbers (Snedecor, [1956], §10.16). If $\mu_1 \neq \mu_2$, mean squares 1 and 4 are used to estimate σ^2 and ρ. There seems to be little justification for using mean square 2 with mean square 4. One or other is biassed or inefficient according as $\mu_1 \neq \mu_2$ or $\mu_1 = \mu_2$.

The bias in any of the above methods is difficult to determine without some knowledge of the effect of culling on the expected values of $\sum_1^2 - [(n - 1)/n]\sigma^2$, $\bar{y}_1 - \mu_1$, $(\bar{y}_1 - \mu_1)^2 - \sigma^2/n$ and $(\bar{y}_1 - \mu_1)\bar{y} - \sigma^2/N$. However, these unknown biasses are clearly undesirable. Since the method of estimation described in this paper is the maximum likelihood method whether or not the culling is random with respect to first records, it is to be preferred. If culling is random with respect to first records, c can be substituted as $c = 1$ in the formulae for the asymptotic bias and variance of the maximum likelihood estimator and the asymptotic efficiency of the regression method.

Wadell [1961] has given a method for estimating ρ when the culling is equivalent to truncation of the first lactation yields. His estimate of ρ is a function of $\bar{y}_1$, $\bar{y}_2$, s^2, Σ_1^2 and the mean squares 2, 3, and 4 of Table 2. The estimate has a negligible bias for large values of n but its variance is not given and therefore its efficiency is unknown.

6. A NUMERICAL EXAMPLE

Dr. A. E. Freeman of the Department of Animal Husbandry, Iowa State University has kindly made available to me some Complete Herd Improvement Registry Records of an Iowa Board of Control herd of Holstein-Friesian cattle. The records included some first and second lactation yields made in various years but expressed as deviations from the herd average for each particular year. The cows were milked twice daily and their yields expressed as mature equivalent 305-day lactation yields.

The values of the various relevant statistics for a sample of the records were:

TABLE 2
ANALYSIS OF VARIANCE OF FIRST AND SECOND RECORDS

Source	d.f	Sums of Squares	Expected Values of Mean Squares	Ref. Nos.
Cows adj. Periods	$N-1$	$\frac{n}{2}(\Sigma_2^2+2\Sigma_{12}-\Sigma_1^2)+Ns^2$	$\sigma^2(1-\rho)+\frac{n+N-2}{N-1}\rho\sigma^2$ $+\frac{n}{2(N-1)}(\rho^2+2\rho-1)\Delta_1$	1
Periods ign. Cows	1	$\frac{nN}{N+n}(\bar{y}-\bar{y}_2)^2$		
Cows and Periods	N	$\frac{n}{2}(\Sigma_2^2+2\Sigma_{12}-\Sigma_1^2)+Ns^2$ $+\frac{nN}{N+n}(\bar{y}-\bar{y}_2)^2$		
Cows ign. Periods	$N-1$	$\frac{n}{2}(\Sigma_2^2+2\Sigma_{12}-\Sigma_1^2)+Ns^2$ $+\frac{nN}{N+n}(\bar{y}-\bar{y}_2)^2-\frac{n}{2}(\bar{y}_1-\bar{y}_2)^2$	$\sigma^2(1-\rho)$ $+\frac{(n+N)^2-(N+3n)}{(N-1)(N+n)}\rho\sigma^2$ $+\frac{n(N-n)}{2(N-1)(N+n)}(\mu_2-\mu_1)^2$ $+\frac{n}{N-1}\Delta_2$	2
Periods adj. Cows	1	$\frac{n}{2}(\bar{y}_1-\bar{y}_2)^2$	$\sigma^2(1-\rho)+\frac{n}{2}(\mu_2-\mu_1)^2+n(1-\rho)\Delta_3$	3
Error	$n-1$	$\frac{n}{2}(\Sigma_2^2-2\Sigma_{12}+\Sigma_1^2)$	$\sigma^2(1-\rho)+\frac{n(1-\rho)^2}{2(n-1)}\Delta_1$	4
TOTAL	$N+n-1$	$n\Sigma_2^2+Ns^2+\frac{Nn}{N+n}(\bar{y}-\bar{y}_2)^2$		
Adj. for mean	1	$\frac{(N\bar{y}+n\bar{y}_2)^2}{N+n}$		

$\Delta_1 = \Sigma_1^2 - \frac{n-1}{n}\sigma^2,$

$$\Delta_2 = \frac{1}{2}(\rho^2+2\rho-1)\,\Delta_1 + \frac{1}{2}\left(\frac{N-n}{N+n}\rho^2+2\rho-1\right)(D^2-\sigma^2/n)$$
$$+\left(\frac{N-n}{N+n}\rho\mu_2+\mu_2-\mu_1+\rho\mu_1\right)D - \frac{2N\rho}{N+n}(D\bar{y}-\sigma^2/N)$$

$\Delta_3 = \frac{1}{2}(1-\rho)(D^2-\sigma^2/n)-(\mu_2-\mu_1)\,D$ and $D = \bar{y}_1-\mu_1$.

$$N = 220, \qquad n = 150, \qquad S = 0.682;$$
$$\bar{y} = -1.118, \qquad \bar{y}_1 = 1.147, \qquad \bar{y}_2 = -0.207;$$
$$s^2 = 286.55, \qquad \Sigma_1^2 = 270.51, \qquad \Sigma_{12} = 125.85 \text{ and } \Sigma_2^2 = 354.98.$$

The maximum likelihood estimator of σ_2^2/σ_1^2 (3.1) is $\widehat{(\sigma_2^2/\sigma_1^2)} = 1.25$. This value is almost certainly not significantly greater than $\sigma_2^2/\sigma_1^2 = 1$. However, it is sufficiently large to raise some doubts about the assumption that $\sigma_2^2 = \sigma_1^2$. The estimate of $\sigma_2^2(1-\rho^2)$ is greater, but not significantly greater, than the estimate of σ_1^2. However, for illustrative purposes we shall assume $\sigma_2^2 = \sigma_1^2$.

Equation (2.5) for $\hat{\rho}$ is

$$\hat{\rho}^3 - 0.3922\hat{\rho}^2 + 5.1257\hat{\rho} - 2.0728 = 0.$$

The simple regression estimator of ρ is

$$b = \Sigma_{12}/\Sigma_1^2 = 0.465.$$

Three cycles of an iteration based on formulae (2.8) and with $\hat{\rho}_0 = b$ as the initial solution show that the value of $\hat{\rho}$ is, to three decimal places,

$$\hat{\rho} = 0.404.$$

This is the only real root of the maximum likelihood equation for $\hat{\rho}$. The two estimators together with their estimated standard errors [(2.7) and (4.1)] are therefore:

$$b = 0.465 \pm 0.084$$

and

$$\hat{\rho} = 0.404 \pm 0.069.$$

The latter standard error is an asymptotic standard error. The estimated efficiency of the regression estimator is only 67 percent and so $\hat{\rho}$ is probably well worth calculating. The estimated asymptotic bias (2.9) is very small,

$$\widehat{E(\delta)} = 0.0004.$$

The maximum likelihood estimator of σ^2 (2.6) is $\hat{\sigma}^2 = 314.46$. This agrees well with the value $s^2 = 286.55$ used above in estimating $E(\delta)$ and the standard errors of b and $\hat{\rho}$.

The estimated value of c is

$$\hat{c} = s^2/\Sigma_1^2 = 1.059.$$

This value of c suggests that there has been very little selection based on first lactation yield in this particular sample of records. The standardized selection differential

$$(\bar{y}_1 - \bar{y})/s = 0.134$$

suggests that the effective selection intensity P was near 0.93.

7. SUMMARY

The maximum likelihood estimation of repeatability from first and second lactation records of a herd subject to culling is discussed. The method is applicable only when it can be assumed that, had there been no culling, the variances of the first and the second lactation records would have been equal, and when all the first records are available whether or not the cow has a second record. The efficiency of the more usual estimate of repeatability, based on the regression of second records on first records, is shown to be low when the repeatability is high. In many cases the calculation of the maximum likelihood estimate would seem to be well worthwhile. The use of the method in estimating heritability is mentioned. An illustrative example is given.

ACKNOWLEDGEMENT

I am grateful to Professor O. Kempthorne for helpful discussions during the preparation of this paper.

REFERENCES

Curnow, R. N. [1957]. Heterogeneous error variances in split-plot experiments. *Biometrika 44*, 378–83.

Finney, D. J. [1957]. The consequences of selection for a variate subject to errors of measurement. *Revue de l'Institut International de Statistique 24*, 22–9.

Henderson, C. R., Kempthorne, O., Searle, S. R. and von Krosigk, C. M. [1959]. The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Lerner, I. M. [1958]. *The Genetic Basis of Selection.* John Wiley and Sons, Inc., New York.

Lush, J. L. [1945]. *Animal Breeding Plans.* 3rd Ed., Iowa State College Press, Ames, Iowa.

Snedecor, G. W. [1956]. *Statistical Methods.* 5th Ed., Iowa State College Press, Ames, Iowa.

Wadell, L. H. [1961]. Selection bias in intraclass correlation repeatability estimates. Unpublished manuscript.

THREE CLASSES OF UNIVARIATE DISCRETE DISTRIBUTIONS

C. G. Khatri and I. R. Patel[1]
M. S. University of Baroda, Baroda, India.

1. INTRODUCTION

Families of descrete distributions have been developed and studied by many authors, including, Neyman [1939], Feller [1943], Skellam [1952], Beall and Rescia [1953] and Gurland [1957, 1958]. These families are of three types:

$$\text{Type } A: \quad g_A(z) = \exp\{h(z)\},$$

$$\text{Type } B: \quad g_B(z) = \{h(z)\}^n,$$

$$\text{Type } C: \quad g_C(z) = c \log\{h(z)\},$$

where $g(z)$ represents a probability generating function (p.g.f.) and $h(z)$ is a p.g.f., except possibly for additive and multiplicative constants. The aim of this paper is to set up formulae for certain statistics for these types. It is hoped that these will be of use to reseach workers in practical fields, who will be formulating compound and generalised distributions of these types by using specific forms of $h(z)$.

2. RECURRENCE RELATIONS FOR PROBABILITIES

Notations: Let the r-th derivation of $f(z)$ be denoted as $f^{(r)}(z)$. Also let

$$P_r = \{g^{(r)}(z)/r!\}\big|_{z=0}, \tag{1}$$

and

$$\pi_r = \{h^{(r)}(z)/r!\}\big|_{z=0}. \tag{2}$$

By the definition of $g(z)$, it is clear that P_r denotes the probability of the r-th count in $g(z)$ and π_r, the probability of the r-th count, excepting possibly for additive and multiplicative constants.

Type A: Here

$$g_A(z) = \exp\{h(z)\}.$$

[1] Present address: Statistical Officer (I), Rajkot, India.

Successive differentiation leads to

$$g_A^{(r)}(z) = \sum_{k=1}^{r} \binom{r-1}{k-1} h^{(k)}(z) g_A^{(r-k)}(z). \tag{3}$$

Hence, on letting $z = 0$, we have

$$P_r = \sum_{k=1}^{r} k\pi_k P_{r-k}/r \quad \text{with} \quad P_0 = \exp(\pi_0). \tag{4}$$

Type B: Here

$$g_B(z) = \{h(z)\}^n.$$

Successive differentiation of $h(z)g_B^{(1)}(z) = ng_B(z)h^{(1)}(z)$ leads to

$$\sum_{k=1}^{r} \binom{r-1}{k-1} h^{(k-1)}(z) g_B^{(r-k+1)}(z) = n \sum_{k=1}^{r} \binom{r-1}{k-1} h^{(k)}(z) g_B^{(r-k)}(z). \tag{5}$$

Hence, on letting $z = 0$, we have

$$P_r = \sum_{k=1}^{r} (nk - r + k)\pi_k P_{r-k}/r\pi_0 \quad \text{with} \quad P_0 = \pi_0^n . \tag{6}$$

Type C: Here

$$g_C(z) = c \log \{h(z)\} \quad \text{where} \quad c = \log \{h(1)\}^{-1}.$$

Successive differentiation of $h(z)g_C^{(1)}(z) = ch^{(1)}(z)$ leads to

$$h(z)g_C^{(r)}(z) = ch^{(r)}(z) - \sum_{k=1}^{r-1} \binom{r-1}{k} h^{(k)}(z) g_C^{(r-k)}(z). \tag{7}$$

Hence, on letting $z = 0$, we have

$$P_r = \{rc\pi_r - \sum_{k=1}^{r-1} (r-k)\pi_k P_{r-k}\}/r\pi_0 \quad \text{with} \quad P_0 = c \log \pi_0 . \tag{8}$$

3. FACTORIAL CUMULANTS

Notations: Let

$$\mu'_{[r]} = \{h^{(r)}(z)\}\,|_{z=1} , \tag{9}$$

$$M'_{[r]} = \{g^{(r)}(z)\}\,|_{z=1} , \tag{10}$$

and

$$K_{[r]} = \{(d/dz)^r \log g(z)\}\,|_{z=1} . \tag{11}$$

From the definition of factorial cumulants, it is clear that $\mu'_{[r]}$ is the r-th factorial moment of $h(z)$ if $h(z)$ is a p.g.f. and $M'_{[r]}$ and $K_{[r]}$ are respectively r-th factorial moment and r-th factorial cumulant of $g(z)$.

Type A: Here $\log \{g_A(z)\} = h(z)$ and hence,

$$K_{[r]} = \mu'_{[r]} \,. \tag{12}$$

Type B: Here

$$\phi(z) = \log g_B(z) = n \log h(z).$$

Hence, on using (11), it is clear that r-th factorial cumulant of $g_B(z)$ is n times the r-th factorial cumulant of $h(z)$ if $h(z)$ is a p.g.f. Also on using (7) (with necessary modifications), (9) and (11), we have

$$K_{[r]} = n\mu'_{[r]} - \sum_{k=1}^{r-1} \binom{r-1}{k} \mu'_{[k]} K_{[r-k]} \,. \tag{13}$$

Type C: Here, it is easy to give a recurrence relation for factorial moments $M'_{[r]}$ rather than factorial cumulants. By using (7), (9) and (10), this relation can be shown to be

$$M'_{[r]} = \left\{ c\mu'_{[r]} - \sum_{k=1}^{r-1} \binom{r-1}{k} \mu'_{[k]} M'_{[r-k]} \right\} / \mu'_0 \,, \tag{14}$$

where $\mu'_0 = h(1)$. The factorial cumulants can be obtained from (14) by using the relations

$$K_{[1]} = M'_{[1]} \,, \qquad K_{[2]} = M'_{[2]} - M'^2_{[1]} \,,$$
$$K_{[3]} = M'_{[3]} - 3M'_{[2]}M'_{[1]} + 2M'^3_1 \,, \qquad \text{etc.}$$

4. SPECIAL CASES

Notation: If a variate has either a Binomial or a Negative Binomial law as a special case, we say that it has a general binomial law.

Type A: From the form of the Poisson-Binomial, Negative-Binomial, generalised Polya Aeppli, Beall and Rescia [1953] and Neyman's Types A, B, and C, it is clear that they belong to this form. It is to be noted, that, if in the classical problem of egg masses and larvae, the egg masses have a Poisson distribution with p.g.f. $\exp [\lambda(z - 1)]$ and the larvae within an egg mass, a distribution with p.g.f. $w(z)$, then the distribution of the larvae over the whole field is $\exp [\lambda\{w(z) - 1\}]$ which is of Type A with $h(z) = \lambda\{w(z) - 1\}$.

Some important distributions with their recurrence relations are:

(i) Poisson-Hypergeometric distribution [1958]: Here

$$w(z) = \sum_{r=0}^{\infty} \binom{k}{r} \alpha_{(r)} m^r (z - 1)^r / (\alpha + \beta)_{(r)} \tag{15}$$

where

$$\alpha_{(r)} = \alpha(\alpha + 1)(\alpha + 2) \cdots (\alpha + r - 1), \qquad \binom{k}{r} = k^{(r)}/r!,$$

$k^{(r)} = k(k-1) \cdots (k-r+1)$, $\alpha_{(0)} = 1$, $k^{(0)} = 1$ and α, β, m and k are such that $w^{(r)}(0)/r! = \pi_r$ is positive. The above distribution was first given by Gurland [1958]. The recurrence relation in probabilities is

$$P_r = \lambda \sum_{s=1}^{r} s\pi_s P_{r-s}/r \tag{16}$$

where $P_0 = \exp\{\lambda(\pi_0 - 1)\}$, and

$$\pi_r = \frac{\alpha + \beta + r - 2 - m(\alpha - k + 2r - 3)}{r(1-m)}\pi_{r-1} - \frac{(\alpha + r - 1)(k - r + 2)m}{r(r-1)(1-m)}\pi_{r-2}\,, \tag{17}$$

$$\pi_0 = \sum_{s=0}^{\infty} \alpha_{(s)}(-k)_{(s)}m^s/s!(\alpha+\beta)_{(s)} \quad \text{and} \quad \pi_1 = -\frac{d\pi_0}{dm}.$$

For the particular cases of the above distribution, we may refer to Gurland [1958]. The recurrence relation for factorial cumulants is

$$K_{[r]} = \lambda\alpha_{(r)}k^{(r)}m^r/(\alpha+\beta)_{(r)}\,. \tag{18}$$

(ii) Poisson-Power series distribution:

$$w(z) = \sum_{i=0}^{\infty} a_i z^i, \tag{19}$$

where a_i's are constants such that $w(z)$ is convergent for some z. The recurrence relation for probabilities is

$$P_r = \lambda \sum_{s=1}^{r} s a_s P_{r-s}/r \quad \text{with} \quad P_0 = \exp\{\lambda(a_0 - 1)\}. \tag{20}$$

The above distribution was first given by Maritz [1952].

Type B: Let the distribution of egg masses be a general binomial with p.g.f. $(1 - p + pz)^n$. Then the distribution of the larvae over the whole field is $[1 - p + pw(z)]^n$, i.e. the Type B distribution with $h(z) = \{1 - p + pw(z)\}$. When $n = 1$, this leads to the distribution with p.g.f. $w(z)$ with an addition (or subtraction) of zeros. If $n > 1$ or $n < 0$, this can be regarded as n-th confluent of $w(z)$.

Some important distributions with their recurrence relations are:

(i) G. Binomial-Hypergeometric distribution: Here $w(z)$ is the same as given in (15). This was first stated by Gurland [1958] in the special case when $n < 0$, $p < 0$. The recurrence relation in probabilities is

$$P_r = p \sum_{s=1}^{r} (ns - r + s)\pi_s P_{r-s}/ra_0\,, \tag{21}$$

where $P_0 = a_0^n$, $a_0 = 1 - p + p\pi_0$ and π_i's are defined in (17).

(ii) G. Binomial-G. Binomial distribution: Here $w(z) = (1 - m + mz)^k$. Hence, $\pi_i = m(k - i + 1)\pi_{i-1}/i(1 - m)$, with $\pi_0 = (1 - m)^k$, $a_0 = 1 - p + p\pi_0$ and $P_0 = a_0^n$. The recurrence formula has the same form as in (21).

(iii) G. Binomial-Poisson distribution: Here $w(z) = \exp\{\lambda(z - 1)\}$. Hence, $\pi_i = \lambda^i\pi_0$, $\pi_0 = \exp(-\lambda)$, $a_0 = 1 - p + p\pi_0$ and $P_0 = a_0^n$. The recurrence formula has the same form as in (21). When $n = 1$ and $p\{1 - \exp(-\lambda)\} = \theta$, this was called by A. C. Cohen [1960] an extension of a truncated Poisson distribution.

Type C: Let the distribution of egg masses be a logarithmic distribution with p.g.f. $\log(1 - \lambda z)/\log(1 - \lambda)$. Then the distribution of a larva over the whole field will be $\log\{1 - \lambda w(z)\}/\log(1 - \lambda)$, i.e. the Type C distribution with $h(z) = 1 - \lambda w(z)$. The important distributions are obtained by considering $w(z)$ as hypergeometric function G. Binomial, Power-series etc.

Choice of a distribution under the condition of no migration:

Let us suppose that the different sites of a colony are distinct and countable, and let there be no migration between sites of a colony. Assuming the same probabilities of arriving at a particular site by any organism, the distribution of r (when r is fixed) organisms in a particular site is Binomial. Now let us suppose that one or more organisms arriving at the colony follow the truncated Negative Binomial law. (This may be true under the wide applicability of the Negative Binomial in biological data, e.g., Bliss [1953], Evans [1953]). Then, it is easy to show that the p.g.f. of the organisms in a particular site (when there is no migration) is

$$1 - \theta + \theta\{(1 + m_1 - m_1z)^{-k_1} - (1 + m_1)^{-k_1}\}/\{1 - (1 + m_1)^{-k_1}\}$$

where $0 < \theta < 1$, $m_1 > 0$, $k_1 > 0$; i.e. $1 - p + p(1 + m_1 - m_1z)^{-k_1}$ for $0 < p < \{1 - (1 + m_1)^{-k_1}\}^{-1}$, $m_1 > 0$ and $k_1 > 0$. Now suppose that the independent results of n sites are combined together. Then the p.g.f. of the distribution of organisms is

$$[1 - p + p(1 + m_1 - m_1z)^{-k_1}]^n$$

which is a particular case of G. Binomial-G. Binomial. The above distribution can be named as Binomial-Negative Binomial.

5. METHODS OF ESTIMATION FOR G. BINOMIAL-G. BINOMIAL

Here the p.g.f. is $[1 - p + p(1 - m + mz)^k]^n$.

Method of moments: Let $T = K_{[4]}/K_{[2]}^2$, $R = K_{[3]}K_{[1]}/K_{[2]}^2$ and

$S = K_{[1]}^2/K_{[2]}$. The approximate value of n is obtained from

$$n^2(TS - 2R^2 + R) + nS(TS - 6R + 6) + S^2(R - 2) = 0.$$

Then k, m and p are estimated from

$$k = \{S^2 + nS + n^2(2 - R)\}/n\{n(1 - R) - S\},$$

$$m = (nK_{[2]} + K_{[1]}^2)/(k - 1)nK_{[1]} \quad \text{and} \quad p = K_{[1]}/knm. \qquad (22)$$

Method of maximum likelihood when n and k are fixed:

Similar to Sprott's results [1958], we have the maximum-likelihood equations as

$$\bar{r} = nk\hat{p}\hat{m} \quad \text{and} \quad L(\hat{m}) = \Sigma a_r G(r) - N = 0 \qquad (23)$$

with $L'(\hat{m}) = (d/d\hat{m})L(\hat{m}) = \sum a_r G(r)[\hat{m}^{-1}\{1 - k(1 - \hat{p})^{-1}\} - \{1 + (1 - \hat{m})[\hat{m}k(1 - \hat{p})]^{-1}\hat{m}^{-1}\bar{r}\Delta\, G(r)\}]$, where $\bar{r}$ is the sample mean, a_r is the frequency at the r-th count, N is the total frequency, $G(r) = (r + 1)\hat{P}_{r+1}/\hat{P}_r\bar{r}$, $\Delta\, G(r) = G(r + 1) - G(r)$ and $\hat{m}$, $\hat{p}$ are maximum likelihood estimates.

From the first approximate estimates $\hat{p}'$, $\hat{m}'$, the new corrected values $\hat{p}''$, $\hat{m}''$ are estimated from

$$\hat{m}'' = \hat{m}' - \{L(\hat{m}')/L'(\hat{m}')\} \quad \text{and} \quad \hat{p}'' = \bar{r}/nk\hat{m}''. \qquad (24)$$

Sample zero frequency when n and k are fixed:

Here the estimates p and m are obtained from

$$a_0 = N[1 - p + p(1 - m)^k]^n \quad \text{and} \quad \bar{r} = nkpm. \qquad (25)$$

It may be noted that when $n = 1$, the two equations in (25) are the same as those in (23).

6. EXAMPLE

In order to illustrate how the above discussion can help an experimenter in the field of curve fitting, we fit here Binomial-Negative Binomial to the data in Distribution 1 of MacGuire *et al.* [1957, Appendix]. From the data, the first four factorial cumulants are $K_{[1]} = 2.5900156$, $K_{[2]} = 0.6877630$, $K_{[3]} = 0.0218497$, $K_{[4]} = 0.9080044$.

Hence, on taking $n = 1$, the solution of k, by the method of moments, correct to first decimal place is -12.0. Then the various estimates of m and p are $m = -0.2204963$, $p = 0.9788582$ by the method of zero-cell frequency or maximum-likelihood and $m = -0.2196584$, $p = 0.982591$ by the method of moments.

The fits are shown in Table 1. The fits of other distributions are given along side for reference purposes only.

TABLE 1.

Count per plot	Obs. fre-quency	Binomial-Neg. Binomial		Negative Binomial [1957]	Poisson; Binomial [1957]	Poisson Power Series[3]
		by M.L. or zero fr.	by method of moments			
0	355	355.000	346.445	324.30	341.84	339.072
1	600	622.478	628.153	660.37	644.37	645.036
2	781	730.994	735.340	734.06	728.03	730.150
3	567	616.306	618.023	610.45	609.14	610.886
4	441	417.545	417.393	408.82	415.60	416.079
5	245	241.296	240.550	236.54	242.72	242.339
6	135	123.567	122.747	122.49	125.17	124.532
7	42	57.406	56.846	58.12	58.20	57.655
8	17	24.632	24.315	25.68	24.76	24.417
9	11	9.395	9.731	10.70	9.75	9.567
10[2]	11	7.301	5.457	4.47	5.42	5.267
χ^2 with ν d.f.		19.184 7.d.f.	20.294 7.d.f.	34.52 8.d.f.	25.52 8.d.f	25.939 8.d.f.

[2] Expected frequencies are from 10 and above

[3] Probabilities are calculated from the p.g.f. $\exp\left\{-(a + b) + az + bz^2\right\}$ where $a = 1.9022526$ and $b = 0.3438815$.

7. ASYMPTOTIC EFFICIENCIES

Here we give the asymptotic efficiencies for the various methods of estimation for m and p only in a G. Binomial-G. Binomial distribution.

The determinant of the information matrix up to order N^{-1} for the maximum likelihood estimates m and p is

$$D_{m,p} = N^2n^2[npk\{k(1 - p) + (1 - m)m^{-1}\}R - \{k(1 - p) - 1\}^2]/(1 - p)^2, \tag{26}$$

where $R = -1 + \sum G^2(r)P_r$ and $G(r) = (r + 1)P_{r+1}/nkpmP_r$.

The determinant of the covariance matrix up to order N^{-1} for the moment estimates of m and p is

$$D = (K_2K_4 + 2K_2^3 - K_3^2)/[Nk(k - 1)nmK_1]^2, \tag{27}$$

where

$$K_1 = k\,n\,m\,p, \qquad K_2/K_1 = 1 + (k - 1)m - (K_1/n),$$

$$K_3/K_1 = 1 + 3(k - 1)m + (k - 1)(k - 2)m^2 - 3(K_2/n) - (K_1^2/n^2)$$

and

$$K_4/K_1 = 1 + 7(k - 1)m + 6(k - 1)(k - 2)m^2 + (k - 1)(k - 2)(k - 3)m^3 - 4(K_3/n) - 6(K_1K_2/n^2) - (K_1^3/n^3) - 3(K_2^2/nK_1).$$

The asymptotic efficiency in the restricted sense is

$$E_1 = 1/D\ D_{m,p}\,. \tag{28}$$

The determinant of the covariance matrix up to order N^{-1} for the sample zero frequency estimates of m and p is

$$D(m, p) = m^2a_0^2\{a_1(1 - P_0) - P_0K_1\}/N^2a_2^2n^2K_1P_0\ , \tag{29}$$

where $a_0 = 1 - p + p(1 - m)^k$, $a_1 = K_2/K_1$, $P_0 = a_0^n$ and $a_2 = 1 - (1 - m)^k - km(1 - m)^{k-1}$. The asymptotic efficiency with respect to the method of moments is

$$E_2 = D/D(m, p). \tag{30}$$

It may be noted that $E_2 = E_1^{-1}$ when $n = 1$.

TABLE 2
Asymptotic Efficiency E_2 Relative to the Method of Moments When $n = 1$ and $k = -1$.

$p\{1 - (1 - m)^k\} = \theta$		m	
	−0.5	−1	−2
0.3	1.492	1.857	2.193
0.6	1.611	2.125	2.630
0.9	2.444	4.000	4.944

8. ACKNOWLEDGEMENT

The authors thank S. K. Katti for his valuable help in preparation of this paper.

REFERENCES.

Beall, G. and Rescia, R. R. [1953]. A generalisation of Neyman's contagious distributions. *Biometrics 9*, 354–86.

Bliss, C. I. [1953]. Fitting a negative binomial distribution to biological data. *Biometrics 9*, 176–200.

Cohen, A. C. [1960]. An extension of a truncated Poisson distribution. *Biometrics 16*, 446–50.

Evans, D. A. [1953]. Experimental evidence concerning contagious distributions in ecology. *Biometrika 40,* 186–221.

Feller, W. [1943]. On a general class of contagious distributions. *Ann. Math. Stat. 14,* 389–400.

Gurland, J. [1958]. A generalized class of contagious distributions. *Biometrics 14,* 229–49.

Gurland, J. [1957]. Some interrelations among compound and generalized distributions. *Biometrika 44,* 265–68.

Maritz, J. S. [1952]. Note on a certain family of discrete distributions. *Biometrika 39,* 196–98.

MacGuire, J. U., Brindley, T. A. and Bancroft, T. A. [1957]. The distribution of European corn borer larvae *Pyrausta Nubilalis* (Hbn.) in field corn. *Biometrics 13,* 65–78.

Neyman, J. [1939]. On a new class of contagious distribution applicable in entomology and bacteriology. *Ann. Math. Stat. 10,* 35–57.

Skellam, J. G. [1952]. Studies in statistical ecology, I. Spatial pattern. *Biometrika 39,* 346–62.

Sprott, D. A. [1958]. The method of maximum likelihood applied to the Poisson-Binomial distribution. *Biometrics 14,* 97–106.

FURTHER CONSIDERATION OF METHODOLOGY IN STUDIES OF PAIN RELIEF

PAUL MEIER
Department of Statistics, University of Chicago, Chicago, Illinois, U.S.A.

AND

SPENCER M. FREE, JR.
Smith, Kline, and French Laboratories, Philadelphia, Pennsylvania, U.S.A.

INTRODUCTION

In an earlier paper [1] we challenged the prevalent view that in comparisons of pain relieving drugs it is always desirable to have "each patient act as his own control," i.e., to test more than one drug on each patient and to estimate treatment contrasts from within-patient differences. We sought data of other investigators to compare with our own, but we were unable to find references which gave sufficient detail to permit investigation of the merits of alternative designs and analyses. It seemed desirable, therefore, to put our own data on record so as to facilitate discussion of the issues raised. This we did, and the response has been gratifying.

This paper presents and discusses some of the suggestions and comments of those who wrote to us. We are particularly indebted to Frederick Mosteller, Department of Statistics, Harvard University, and to Robert Curnow, A. R. C. Unit of Statistics, University of Aberdeen, Aberdeen, Scotland, who, in addition to making numerous thoughtful comments, performed their own analyses of our data.

A SIMPLE EXTENSION OF THE MODEL

As stated in our earlier paper, this study of relief of post-operative pain was designed as an incomplete block experiment. Each patient was given a dose of a test drug, and the dose was repeated when the patient again complained of severe pain. The number of hours of "greater than 50 percent pain relief" cumulated over both intervals was the measure of drug efficacy. A second drug was tested on each patient in the same way, starting at the time when the patient again complained of severe pain. Three drugs, two levels of a new drug and one level of Demerol, were under study.

It was found that the total hours of relief achieved when a given drug was the second administered was larger on the average than when that drug was administered first. This is consistent with the presumption that the pain decreases with time after operation. However, it invalidates a straightforward least squares analysis based on the model

$$\text{hours of relief} = \text{grand mean} + \text{patient effect} + \text{drug effect} + \text{random error}.$$

We considered the possibility of extending the analysis to include a simple time-period effect, representing the average differences in relief between the second and first periods, but for two reasons we did not pursue it. Firstly, and most importantly, it seemed to us that such a model would be unlikely to be an accurate reflection of the true situation. For example, if the duration of drug effect increases with time after operation, the period effect should be greater when the second drug is given after a potent drug than when it is given after a weak drug. Secondly, our object in the first paper was to compare *simple* alternatives for the design and analysis of studies to evaluate analgesics, and a procedure requiring adjustment for time period as well as patient effects did not seem attractive as a procedure to recommend to clinicians for routine screening of new drugs. (In this connection one should remember that it is almost always necessary to discontinue study of some patients, often for reasons unrelated to drug response. Thus, balance in design is hardly ever achieved.)

Curnow extended the least squares analysis of our data to take account of such a time-period effect. The model becomes

$$\text{hours of relief} = \text{grand mean} + \text{patient effect} + \text{drug effect} + \textit{time period effect} + \text{random error}$$

and the corresponding analysis is given in Table I. He stated that the differences due to the drug which follows a potent one being administered rather later than one which follows a weak one were small and had little effect on the estimates of drug differences. Thus the internal evidence of the experiment itself does not seem to confirm the fears expressed above, and the least squares analysis that includes a time-period effect appears to yield a reasonable description of the data. In particular, in both the inter- and intra-block analyses, the between-group residual mean square is now not larger than the corresponding within-groups mean square.

Curnow went on to point out that the inter-block error is not much larger than the intra-block error, and that one might on this ground justify an analysis *ignoring patient differences*, as shown in Table II.

TABLE I

Least Squares Intra-Block Analysis

Source	D.F.	S.S.	M.S.
Blocks	42	607.256	
Periods and Drugs	3	378.015	
Periods ignoring drugs	1	261.628	
Drugs adjusted for periods	2	116.387	58.194
Between Groups Residual	3	3.051	1.017
Within Groups Residual	37	428.945	11.593
Total	85	1417.256	

Drug Comparisons

Comparison	Mean Difference	Variance	t
D vs. T_1	3.421	1.178	3.15
D vs. T_3	2.037	1.061	1.98
T_3 vs. T_1	1.384	1.025	1.35

TABLE II

Least Squares Analysis Ignoring Patient Differences

Source	D.F.	S.S.	M.S.
Periods and Drugs	3	434.368	
Periods ignoring drugs	1	261.588	
Drugs adjusted for periods	2	172.780	86.390
Remainder	82	982.888	11.986
Total	85	1417.256	

Drug Comparisons

Comparison	Mean Difference	Variance	t
D vs. T_1	3.548	0.874	3.80
D vs. T_3	1.861	0.831	2.04
T_3 vs. T_1	1.687	0.816	1.87

This observation suggests that in future experiments conducted under similar conditions we should *not* arrange matters to eliminate patient effects, if it will cost us much to do so.

Our own calculations and scatter diagrams confirm Curnow's results. However, in view of our first argument against the use of a simple time-period effect we found them surprising and made some further investigation. We take note first that the fact that a model "fits", in the sense that certain internal checks are satisfied, does not insure its correctness. If, for example, second-dose relief were more variable than first-dose relief, an examination of residuals would not in this case establish that fact; the estimated first-dose residual for a given patient has precisely the same magnitude as the estimated second-dose residual for that patient when we use the present model. The second-dose relief scores are, in fact, more variable than are those for first doses. This may be observed by examination of Table 1b in [1] and it is reflected in the fact that the *within patient* residual variance (Table 3a), which excludes the component of variability due to patient differences, is actually larger than the *between patient* variance for first dose relief (Table 2).

If the facts are in accord with the assumption that the increment in relief with increasing time after administration is proportional to time after operation, the effect of the model having only a simple time-period effect would be to assign a part of the drug difference to the time-period effect and thus to reduce the measured differences between drugs, as compared to an analysis for the first period only. In fact, the difference between the most and least potent drugs (D and T_1) estimated by use of the extended model is less by about 15 percent than the estimate obtained from analysis of the first period only.

Granted, then, that the model with a simple time-period effect fits the data insofar as intra-block checks are concerned, what may we expect to gain or lose if we go ahead and use it? We do have some evidence which tends to contradict the assumptions of this model, but the effects do not seem large. We have some slight evidence that we may lose discriminating power by misinterpreting drug differences as time-period effects, but, as judged from these data, the reduction in variance may well compensate. Insofar as testing (i.e., deciding which drug is better) is concerned, or estimating potency in an experiment with a full range of standards, the analysis based on this model should give essentially valid results.

In any event, apart from the question of the best analysis for these data, the smallness of the between patients component in the extended analysis suggests that we would have achieved at least equal precision for drug comparisons if we had kept each patient on a single drug throughout the experimental period. Had we done so, we would have had, in addition to possible benefits in precision, the comfort of closely imitating the ordinary clinical situation, and the advantage of an unbiased estimate of a clearly interpretable measure of effectiveness along with an unbiased estimate of its variance. We might also have had the cooperation of several more surgeons, some of whom, understandably, object to studies which require administration of several coded drugs to each patient.

A REGRESSION MODEL

In furtherance of the point of view that the cost of statistical work is generally small compared to the cost of gathering clinical data, Mosteller suggested that the data be analyzed in accordance with a model which might give a fairly realistic appraisal of the time-trend effect. In particular, he suggested that we not cumulate the two doses of each drug but analyze the data for single doses according to the following model.

$$y_{imd} = \alpha_i + \beta_m + \gamma_m t_{imd} + \epsilon_{imd} ,$$

where

α_i = effect of i-th subject,
β_m = effect of medication m when given at time zero,
γ_m = regression of effect of medication m on time of administration,
t_{imd} = time at which individual i receives the d-th dose of medication m,
ϵ_{imd} = random error.

This analysis was carried out, using the times of drug administration (not shown, but available upon request) in addition to the data already presented. It was found that the regression coefficients γ_m were quite close to one another and, if we restrict the model to the case of equal γ_m , the least squares estimate of this common value is 0.20, corresponding to an additional hour of drug relief for every five hour interval between operation and drug administration. Using this restricted model and doubling the estimated effects to make these results comparable to those for our other analyses, we have the estimates of treatment differences shown in Table III.

Assuming for now the correctness of the regression model, we see that these estimates are quite close to those obtained in the analysis using only the first drug for each patient—Table 2 of [1]. However, the variances are reduced by about one-third, so that the regression

TABLE III

Drug Comparisons from Regression Model Analysis

Comparison	Mean Difference	Variance	t
D vs. T_1	4.066	0.824	4.48
D vs. T_3	2.530	0.732	2.96
T_3 vs. T_1	1.534	0.708	1.82

analysis, using all the data, appears to have about 50 percent greater precision than the "first drug only" analysis, which uses only half the data.

Comparing the apparent precisions provided by the above analyses, we see that Curnow's intra-block analysis, adjusting for patients and time periods (Table I) appears comparable in precision to the "first drug only" analysis. Curnow's analysis ignoring patient effects (Table II) appears almost equal in precision to the analysis based on the regression model.

OTHER COMMENTS AND ERRATA

A third correspondent, Irwin Bross, Roswell Park Memorial Institute, Roswell Park, New York, raised several points, in part overlapping those above. In addition he pointed out that the results of the "least squares" analysis (Table 4a in [1]) differ appreciably from the analysis based on simple linear combinations of intra-individual contrasts. Since the design is nearly balanced, closer agreement might be expected, even though our "least squares" analysis fails to allow for time-period effects. In fact, contrary to the suggestion in [1], the "linear combinations" analysis is not really comparable to the "least squares" analysis. The reason is that in estimating, say, $D - T_1$, no account was taken of the information about this difference given by contrasting the direct estimate of $D - T_3$ with the estimate of $T_1 - T_3$. If we take an average of the directly observed difference, $D - T_1$, with the contrast obtained by combining $D - T_3$ with $T_1 - T_3$, weighting inversely as the estimated variances, and proceed similarly for the other comparisons, we get Table IV, which agrees much more closely with the "least squares" analysis (Table 4a in [1]). As might be expected, the "extended least squares" analysis which does allow for time-period effects (Table I) gives results in excellent agreement with Table IV.

This revision leads in turn to a more precise combined intra- and inter-block analysis in place of that shown in Table 3c in [1]. The result (not given) is quite close to the "least squares" combined analysis.

TABLE IV

DRUG COMPARISONS FROM ANALYSIS OF WITHIN PATIENT CONTRASTS

Comparison	Mean Difference	Variance	t
D vs. T_1	3.449	1.206	3.18
D vs. T_3	2.051	1.073	1.98
T_3 vs. T_1	1.398	1.031	1.38

Both Curnow and Bross pointed out that erroneous entries are shown in Table 4c in [1]. This table should be replaced with Table 4c (revised) as shown.

TABLE 4c (revised)

COMBINED INTRA- AND INTER-BLOCK ANALYSIS FOR DRUG DIFFERENCES

$$w = \frac{1}{1.174} = 0.8518 \qquad w' = \frac{1}{3.720} = 0.2688$$

Calculation of Average Mean Difference: Drug D vs. Drug T_1

$$\frac{(w)(3.150) + (w')(3.900)}{w + w'} = 3.330.$$

Calculation of Variance of Difference

$$\frac{1}{w + w'} = \frac{1}{0.8518 + 0.2688} = 0.892.$$

DRUG COMPARISONS

Comparison	Mean Difference	Variance	t
D vs. T_1	3.330	0.892	3.52
D vs. T_3	1.843	0.848	2.00
T_3 vs. T_1	1.442	0.828	1.58

DISCUSSION

The point which we wished to emphasize in [1] was that on account of the increase in duration of relief with time of administration, the use of simple intra-patient contrasts—"each patient his own control"—may not be optimal, and that an experiment in which each patient received only a single drug might be preferable.

The analyses proposed by Curnow and Mosteller have demonstrated that for this data a model taking account explicitly of the effect of time of drug administration will yield more efficient estimates than does the one-way analysis of the first period only. With respect to the analysis of the present data, both the Curnow and Mosteller models appear to fit quite well, and the estimates of the drug effects produced by their analyses are more precise than those derived from the analysis for the first period only.

It is not clear, however, whether the above finding should be construed as evidence supporting the need for more complex models and analyses than are currently in vogue, or whether instead it is evidence supporting our original viewpoint. If, as the evidence of Curnow's analysis suggests, patient differences are negligible, we would have no reason at all to use each patient as his own control. Were each patient given only one drug, the time-period effect of Curnow's model would be eliminated from treatment contrasts and the one-way analysis—using now the whole of the data—should be fully efficient.

Even if patient differences are not negligible, the advantages of simplicity, both of design and analysis, may outweigh a small gain in efficiency which could be achieved with the use of a more complex design and analysis. With our data, for example, the estimates of greatest apparent precision were those provided by the regression model, and the gain in precision compared to the analysis of the first period only was 50 percent. Such a gain is, of course, quite worth having, but it is not overwhelming, and it is easy to believe that had our design assigned one drug to each patient throughout the period of study, the one-way analysis might have equalled the present regression analysis in precision.

In conclusion, we must again emphasize that we do not claim that our findings apply to all kinds of pain studies. There may be many situations for which within-patient contrasts are far superior to between-patient contrasts. What we do claim is that no principal such as "each patient his own control" is entitled to the status of dogma. In some situations, at least, the simpler methods are better.

REFERENCES

[1] Meier, Paul, Free, S. M., and Jackson, G. L. [1958]. Reconsideration of methodology in studies of pain relief. *Biometrics 14*, 330–42.

FITTING A GEOMETRIC PROGRESSION TO FREQUENCIES

E. J. Williams
Division of Mathematical Statistics,
C. S. I. R. O., Canberra, Australia

SUMMARY

This paper discusses the interpretation of frequency data when the series of observed frequencies is assumed to arise from a population in which the expected frequencies form a geometric progression. Such situations occur in the study of steadily increasing insect populations and similar phenomena, where the common ratio of frequencies is related to the rate of growth of the population.

The estimation of the common ratio, and tests for the significance of departure from geometric trend, are discussed. Asymptotic formulae for the tail probabilities in large samples are determined.

The methods and tests of significance are illustrated by application to some experimental data.

I. INTRODUCTION

Observations are often recorded by classifying them into several classes and counting the frequencies of occurrence in each class. In general, the expected frequencies will be partly specified by theoretical considerations, but will often also depend on one or more unknown parameters.

The interpretation of such data then involves, firstly, testing its concordance with the assumed form of expected frequencies, and secondly, estimating the unknown parameters. Thus, for instance, in a steadily increasing population of organisms, the number of individuals expected in successive age-groups will be in geometric progression, the common ratio between the expected numbers representing not only the growth of the population but also the effects of mortality, migration and other factors. In such a study, one of the objects would be to test the concordance with the assumed geometric progression, and the other would be to estimate the common ratio accurately.

In general problems of this kind, the parameters may be estimated, and the accuracy of the estimates assessed, by the method of maximum likelihood. To test the fit of the model to the data, the χ^2 test is generally appropriate. If hypothetical values of the parameters have been specified, as occurs in many practical instances, their concordance with the estimates derived from the data can also be tested by means of χ^2.

In the particular case, with which this paper deals, of expected frequencies in geometric progression, there exist sufficient statistics for the two parameters involved, representing the general level of the frequencies and their geometric trend. This facilitates estimation and significance testing, although, since the equations of estimation are in general non-linear, iterative methods are required in the arithmetical work of calculating estimates.

The need to fit frequencies in this way arose from a method, devised by Dr. R. D. Hughes, Division of Entomology, C.S.I.R.O., of using the age distribution in an aphid population to study its rate of growth in the field. The device of using the proportions of individuals in the immature instars in a field population of aphids to give, firstly, the age distribution, and secondly, the rate of increase of the population, is the subject of a forthcoming paper by Hughes. In the present paper, some preliminary observations of instar distribution under controlled experimental conditions are discussed.

Chapman and Robson [1960] have considered the age distribution in a stationary population subject to constant mortality. As this formulation leads to a geometric progression of expected frequencies, many of their results anticipate results given in this paper. However, the general objects and scope of the two papers are different.

Aphid populations under uniform and favourable conditions increase at a constant rate, so that their expected numbers at equal intervals of time form a geometric progression. However, the numbers even of an initially small population quickly become too large to be counted accurately. An estimate of the rate of growth of the population may then be made from the age distribution. The immature aphid passes through four instars before reaching maturity. In a stationary population, the number expected in each instar is proportional to its duration. For some species of aphids, for instance *A. craccivora*, the average durations of the first three instars are probably equal under constant conditions, so that equal numbers will be expected in each instar in a stationary population. When the population is increasing at a constant rate, the numbers in each instar will approximate to a geometric progression.

If the common duration of each of the first three instars is c, and the growth-rate of the population is ρ, then the ratio of the number in each instar to that in the preceding one is approximately

$$e^{-c\rho}.$$

In this result, mortality has been neglected, since it has been shown that under favourable conditions mortality in the immature stages is negligible.

By taking a random sample from the aphid population and determining the number in each instar we can first check the validity of the assumption of uniform growth by testing whether the numbers are consistent with a geometric progression. Having satisfied ourselves on this point, we can then estimate the common ratio of the numbers. If c, the duration of each instar, is known, the growth-rate can then be determined.

If the durations of the different instars are different, the above method will be modified, and the estimation of the growth-rate is a little more complicated; in principle, however, such data will still provide information from which the growth-rate can be determined.

II. THE BASIC DISTRIBUTION

We consider a sample of N individuals classified into k classes, the observed frequency in class i being n_i, with expected value

$$E(n_i) = \lambda\mu^i \qquad (i = 0, 1, \cdots, k - 1).$$

If N is assumed to have a Poisson distribution, so has each of the n_i; the joint probability density is then

$$P(n_0, n_1, \cdots, n_{k-1}) = \prod (e^{-\lambda\mu^i}\lambda^{n_i}\mu^{in_i}/n_i!) = \frac{e^{-\lambda\phi_0(\mu)}\lambda^N\mu^T}{n_0!\,n_1!\cdots n_{k-1}!}, \quad (1)$$

where

$$T = n_1 + 2n_2 + \cdots + (k - 1)n_{k-1},$$

and

$$\phi_0(\mu) = 1 + \mu + \cdots + \mu^{k-1}.$$

From the form of the density it is apparent that N and T are a pair of sufficient statistics for the parameters λ and μ. Thus, questions of estimation may be referred to the joint distribution of N and T.

By summing the probability (1) over values of the n_i leading to the given values of N and T, we find the joint distribution of N and T as

$$P(N, T) = \frac{C(N, T)}{N!}\, e^{-\lambda\phi_0(\mu)}\lambda^N\mu^T, \quad (2)$$

where $C(N, T)$ is a numerical coefficient.

The number N has a Poisson distribution with mean $\lambda\phi_0(\mu)$;

$$P(N) = e^{-\lambda\phi_0(\mu)}[\lambda\phi_0(\mu)]^N/N!.$$

Hence for given N, the conditional density of T, which is independent of λ, is

$$P(T \mid N) = C(N, T)\mu^T/[\phi_0(\mu)]^N.$$

This is a special case of the distribution discussed by Noack [1950] and described by him as a power-series distribution. Our $[\phi(\mu)]^N$ takes the place of his $f(z)$. From this expression for the density we see that $C(N, T)$ is the coefficient of μ^T in $\phi_0(\mu)^N$. This fact enables the numerical coefficient to be determined directly.

This conditional density of T provides the basis for estimates and tests of significance about μ.

III. THE COMBINATORIAL FACTOR

The coefficient $C(N, T)$ is seen to be a generalization of the binomial coefficient (for the positive binomial, when $k = 2$; for the negative binomial, when $k = \infty$). It is the number of ways of allocating T like objects among N different cells, none of which may contain more than $k - 1$ objects (see Riordan [1958], page 104, where some recurrence relations and moment formulae are also given).

The coefficients are generated by the function

$$[\phi_0(\mu)]^N = \left(\frac{1 - \mu^k}{1 - \mu}\right)^N = \left[\sum \binom{N + i - 1}{N - 1}\mu^i\right]\left[\sum \binom{N}{j}(-\mu^k)^i\right].$$

On equating coefficients of μ^T we find

$$C(N, T) = \binom{N + T - 1}{N - 1} - \binom{N}{1}\binom{N + T - k - 1}{N - 1} + \binom{N}{2}\binom{N + T - 2k - 1}{N - 1} - \cdots \qquad (3)$$

The series has $1 + [T/k]$ terms. This expansion in terms of binomial coefficients is useful, especially if k is large, since the first few terms then give a close approximation.

Because of symmetry, $C(N, T) = C[N, (k - 1)N - T]$ so that results for $T > \frac{1}{2}(k - 1)N$ can be found from those for $T < \frac{1}{2}(k - 1)N$.

From the generating function or otherwise we may also deduce the recurrence relation

$$C(N, T) = C(N, T - 1) + C(N - 1, T) - C(N - 1, T - k)$$

which is useful for computation.

Hitherto we have been considering relations among the $C(N, T)$ for a fixed value of k. We now consider relations for different k, and shall indicate by a suffix the number of classes.

We may express the coefficients for k classes in terms of the coefficients corresponding to the factors of k, by means of the following reduction formula.

If $k = fg$,

$$\phi_{0k}(\mu) = \frac{1 - \mu^k}{1 - \mu} = \frac{1 - \mu^f}{1 - \mu} \cdot \frac{1 - \mu^{fg}}{1 - \mu^f} = \phi_{0f}(\mu) \cdot \phi_{0g}(\mu^f).$$

Hence

$$C_k(N, T) = C_f(N, T) + C_f(N, T - f)C_g(N, 1) + C_f(N, T - 2f)C_g(N, 2) + \cdots ,$$

a series of $1 + [T/f]$ terms. The terms of these series are all positive, and generally smaller than those of the series (3) of products of binomial coefficients. They may have some advantages for computation, provided the coefficients for f and g classes are known.

By differentiating the generating function with respect to its parameter we may prove recurrence formulae of the type

$$\begin{aligned} C(N, T) = {} & \frac{N}{(k - 1)N - 2T} \\ & \cdot [(k - 1)C(N - 1, T) + (k - 3)C(N - 1, T - 1) \\ & + \cdots - (k - 1)C(N - 1, T - k + 1)]. \end{aligned} \tag{4}$$

When $k = 3$, many particularly simple recurrence formulae may be established. Recurrence relations when $k = 3$ are

$$\begin{aligned} C(N, T) &= \frac{N}{T(2N - T)} \\ &\quad \cdot [(2N - 1)C(N - 1, T - 1) + 3(N - 1)C(N - 2, T - 2)] \\ &= \frac{1}{2T} [(2N - T + 1)C(N, T - 1) + 3NC(N - 1, T - 2)] \\ &= \frac{N}{N - T} [C(N - 1, T) - C(N - 1, T - 2)], \end{aligned}$$

the last being a particular case of (4). As k increases the recurrence formulae become more complicated.

IV. MOMENTS OF THE CONDITIONAL DISTRIBUTION OF T

The moments and cumulants of the conditional distribution of T are of general interest, and will also be of use when the determination of probabilities in the tails of the distribution is being considered.

Since the cumulants for a sample of N are simply N times those for a sample of 1, we shall consider a sample of 1, for which T equals X, the number of the class $(0, 1, \cdots, k - 1)$ in which the observation falls.

Clearly, if

$$\phi_r(\mu) = \left(\mu \frac{d}{d\mu}\right)^r \phi_0(\mu) = \sum_{x=0}^{k-1} x^r \mu^x,$$

then the rth moment about zero is $\phi_r(\mu)/\phi_0(\mu)$. This result is equivalent to the results of Noack [1950], though expressed in a slightly different manner.

However, unless k is small, these expressions are not convenient for the computation of the central moments and cumulants, which are more simply found directly,

The moment-generating function of X is

$$E(e^{sX}) = \sum \mu^x e^{sx} / \sum \mu^x = \phi_0(\mu e^s)/\phi_0(\mu).$$

Thus the cumulant-generating function is

$$\begin{aligned} K(s) &= \log \phi_0(\mu e^s) - \log \phi_0(\mu) \\ &= -\log (1 - \mu e^s) + \log (1 - \mu^k e^{sk}) + \log (1 - \mu) - \log (1 - \mu^k) \\ &= \sum \frac{\mu^r}{r} (e^{rs} - 1) - \sum \frac{\mu^{rk}}{r} (e^{rsk} - 1). \end{aligned}$$

From this expansion we derive the particular results

$$\kappa_1 = \frac{\mu}{1 - \mu} - \frac{k\mu^k}{1 - \mu^k},$$

$$\kappa_2 = \frac{\mu}{(1 - \mu)^2} - \frac{k^2 \mu^k}{(1 - \mu^k)^2},$$

$$\kappa_3 = \frac{\mu + \mu^2}{(1 - \mu)^3} - \frac{k^3(\mu^k + \mu^{2k})}{(1 - \mu^k)^3},$$

$$\kappa_4 = \frac{\mu + 4\mu^2 + \mu^3}{(1 - \mu)^4} - \frac{k^4(\mu^k + 4\mu^{2k} + \mu^{3k})}{(1 - \mu^k)^4}.$$

The form of the cumulant-generating function shows that, if u_k is a geometric variable with parameter μ^k, then $X + ku_k = u_1$. It also follows that, if $u_k(N)$ is a negative binomial variable with parameter μ^k and index N, then $T + ku_k(N) = u_1(N)$.

When $\mu = 1$, the distribution is the uniform discrete distribution, and the cumulants are expressible in terms of Bernoulli's numbers. We then have

$$K(s) = -\log\left(\frac{e^s - 1}{s}\right) + \log\left(\frac{e^{sk} - 1}{sk}\right)$$

$$= \frac{k-1}{2}s + (k^2 - 1)\frac{B_2}{2}\frac{s^2}{2!} + (k^4 - 1)\frac{B_4}{4}\frac{s^4}{4!} + \cdots,$$

where $s/(e^s - 1) = 1 + \sum B_r s^r/r! = e^{Bs}$ in symbolic form. Hence $\kappa_r = (k^r - 1)B_r/r$; in particular, $\kappa_1 = (k - 1)/2$, $\kappa_2 = (k^2 - 1)/12$, $\kappa_3 = 0$, $\kappa_4 = -(k^4 - 1)/120$.

Relation between Cumulants Corresponding to μ and μ^{-1}.

When the series of frequencies is reversed, the ratio μ is replaced by its reciprocal. It therefore follows that, when μ is replaced by μ^{-1}, the odd cumulants other than κ_1 are simply changed in sign, and the even cumulants are unaffected. We have

$$\kappa_1(\mu^{-1}) = k - 1 - \kappa_1(\mu),$$
$$\kappa_r(\mu^{-1}) = (-1)^r\kappa_r(\mu) \qquad (r > 1).$$

These results may also be readily verified from the form of the cumulant-generating function.

Because of these facts, we need not consider values of μ outside the range (0, 1).

V. ESTIMATION OF THE RATIO OF FREQUENCIES

In the conditional distribution, T is a sufficient statistic for the parameter μ, the common ratio of the expected ferquencies. Thus the estimation of μ is straightforward. However, since the equation of estimation by the method of maximum likelihood is non-linear, iterative methods will be required to solve it.

The logarithm of the conditional probability of T, apart from terms independent of the parameter, is $L = T \log \mu - N \log \phi_0(\mu)$, and its first derivative with respect to μ is

$$(T/\mu) - (N\phi_1/\mu\phi_0). \tag{5}$$

Equating the derivative to zero gives the maximum likelihood estimator of μ. We indicate by an asterisk the maximum likelihood estimator and functions of it. Thus

$$T - (N\phi_1^*/\phi_0^*) = 0. \tag{6}$$

Since the equation is linear in T, it is clear that the estimator is to be found simply by equating T to its expectation: $T = N\kappa_1^*$, as may be verified from the results of the previous section.

The variance of the derivative (5) gives the information I_μ about μ in the sample. Being linear in T, the derivative has variance

$$V(T)/\mu^2 = N\kappa_2/\mu^2 = \frac{N}{\mu^2}\left(\frac{\mu}{(1-\mu)^2} - \frac{k^2\mu^k}{(1-\mu^k)^2}\right) = I_\mu .$$

This result is a special case of the results of Patil [1961], who investigates the estimation of the parameter of a generalized power-series distribution. In large samples, the reciprocal of I_μ approximates the variance of the estimate μ^*; that is, $V(\mu^*) = \mu^2/N\kappa_2$.

For purposes of calculating and tabulating the solutions of the maximum likelihood equation, we shall put

$$T/(k-1)N = v,$$

so that, for all k, $0 \leq v \leq 1$.

Then the estimating equation may be written

$$\kappa_1 = \frac{\mu}{1-\mu} - \frac{k\mu^k}{1-\mu^k} = (k-1)v. \tag{7}$$

An alternative form is

$$\frac{1}{1-\mu} - \frac{k}{1-\mu^k} = -(k-1)(1-v),$$

or

$$\frac{\mu^{-1}}{1-\mu^{-1}} - \frac{k\mu^{-k}}{1-\mu^{-k}} = (k-1)(1-v).$$

Thus, if μ^* is the root corresponding to v, then μ^{*-1} is the root corresponding to $1 - v$. In particular, if $v = \frac{1}{2}$, $\mu^* = 1$.

Since the roots corresponding to values of v exceeding $\frac{1}{2}$ are the reciprocals of roots corresponding to values of v less than $\frac{1}{2}$, we may henceforth confine attention to $v \leq \frac{1}{2}$, $\mu \leq 1$.

Equation (7) can be solved iteratively, once an approximate value of μ has been chosen. If v is not too near $\frac{1}{2}$, and k is not small, a first approximation is

$$\mu_1 = \frac{(k-1)v}{1+(k-1)v} = \frac{T}{T+N}, \tag{8}$$

and a second approximation is

$$\mu_2 = \mu_1 + k\mu_1^k(1-\mu_1)^2.$$

The difference between the two sides of (7) when an approximate value of μ is substituted represents $-\mu/N$ times (5), the first derivative of the likelihood. The adjustment to μ is given by the ratio of this difference to $-\mu I_\mu/N$. Then, approximately,

$$\mu^* = \mu\left[1 - \frac{\kappa_1 - (k-1)v}{\kappa_2}\right].$$

The substitution and adjustment may be repeated as often as required to give the desired accuracy.

If v is close to $\frac{1}{2}$, an approximation alternative to (8) is to be preferred. We put

$$\mu = e^{-\theta}, \qquad v = \tfrac{1}{2} - w.$$

Then equation (7) becomes

$$\frac{e^{-\theta}}{1 - e^{-\theta}} - \frac{ke^{-k\theta}}{1 - e^{-k\theta}} = (k-1)(\tfrac{1}{2} - w)$$

or, symbolically in terms of Bernoulli's numbers,

$$\frac{1}{\theta}(e^{B\theta} - e^{Bk\theta}) = (k-1)(\tfrac{1}{2} - w).$$

We then find

$$\frac{B_2\theta}{2!}(k^2 - 1) + \frac{B_4\theta^3}{4!}(k^4 - 1) + O(\theta^5) = (k-1)w,$$

whence

$$\theta = \frac{12w}{k+1} + \frac{144}{5}\frac{(k^2+1)w^3}{(k+1)^3} + O(w^5),$$

and

$$\mu^* = 1 - \frac{12w}{k+1} + \frac{72w^2}{(k+1)^2} - \frac{144}{5}\frac{(k^2+11)w^3}{(k+1)^3} + O(w^4). \qquad (9)$$

Solutions of the maximum likelihood equation for various values of k and v are given in Table 1. Once values of μ are given, it is easy to compute I_μ. As we shall see in Section X, the solution not only gives a point estimate of the ratio μ, and an approximate standard error based on the information function I_μ, but also gives a means of determining tail probabilities, needed in making significance tests and setting confidence limits.

VI. ESTIMATION OF λ

In general, for the problems considered in this paper, the actual value of λ, representing the size of the population (or rather, of the

TABLE 1

Maximum Likelihood Estimate of μ [with Argument $v = T/(k-1)N$]

k	2	3	4	5
0.00	0	0	0	0
0.05	052632	092894	131334	167119
0.10	111111	178395	238358	291010
0.15	176471	261940	333333	392939
0.20	250000	346500	422530	483323
0.25	333333	434259	509668	567737
0.30	428571	527202	597388	649654
0.35	538462	627431	687922	731617
0.40	666667	737405	783468	815815
0.45	818182	860221	886484	904443
0.50	1.000000	1.000000	1.000000	1.000000

zero-class), is not of interest. However, for testing the significance of departure of the frequencies from a geometric progression, estimates of the expected frequencies in each class will sometimes be required. We shall then be interested in the simultaneous estimation of λ and μ, based on N and T. We therefore consider this question, rather than the estimation of λ alone.

The joint probability of N and T, as given in (2), is

$$\frac{C(N, T)}{N!} e^{-\lambda\phi_0(\mu)}\lambda^N\mu^T,$$

so that the logarithm of the likelihood, considered as a function of the two parameters, and with constant factors ignored, is $L = -\lambda\phi_0(\mu) + N \log \lambda + T \log \mu$. The derivatives are

$$\begin{aligned} L_\lambda &= -\phi_0(\mu) + (N/\lambda), \\ L_\mu &= [-\lambda\phi_1(\mu) + T]/\mu. \end{aligned} \tag{10}$$

The estimates are given when these derivatives are equated to zero, or when N and T are equated to their expected values. For direct solution, λ can be eliminated between the two equations, giving an equation for μ identical with (6). μ having been found, λ may then be found by substitution in (10).

For simultaneous solution, and also to give a measure of the information about λ and μ, we require the covariance matrix of the derivatives. Noting that N is a Poisson variate, with variance equal to its mean $\lambda\phi_0(\mu)$, that the covariance of N and T is $\lambda\phi_1(\mu)$ and that the variance of T is $\lambda\phi_2(\mu)$, we find for the covariance matrix of the derivatives

$$\mathbf{I} = \lambda\begin{bmatrix} \phi_0/\lambda^2 & \phi_1/\lambda\mu \\ \phi_1/\lambda\mu & \phi_2/\mu^2 \end{bmatrix}.$$

The inverse of **I**, giving approximate variances and covariance of the estimates in large samples, is

$$\mathbf{I}^{-1} = \frac{1}{\lambda(\phi_0\phi_2 - \phi_1^2)}\begin{bmatrix} \lambda^2\phi_2 & -\lambda\mu\phi_1 \\ -\lambda\mu\phi_1 & \mu^2\phi_0 \end{bmatrix}.$$

From trial values λ, μ of the parameters we find improved estimates by means of the equations

$$\begin{bmatrix} \lambda^* \\ \mu^* \end{bmatrix} = \begin{bmatrix} \lambda \\ \mu \end{bmatrix} + \mathbf{I}^{-1}\begin{bmatrix} L_\lambda \\ L_\mu \end{bmatrix},$$

which give explicitly

$$\begin{aligned} \lambda^* &= \frac{N\phi_2 - T\phi_1}{\phi_0\phi_2 - \phi_1^2}, \\ \mu^* &= \mu\left[1 - \frac{N\phi_1 - T\phi_0}{\lambda(\phi_0\phi_2 - \phi_1^2)}\right]. \end{aligned} \qquad (11)$$

Note that the trial value of λ does not appear in the equation for λ^*, since the estimating equations are linear in λ^*. The adjustments given by (11) may be repeated to give the accuracy required.

The expected frequency in class i is $\lambda\mu^i$; the variance of the estimated frequency may be determined, using $\mathbf{I}^{-1}$, as

$$\frac{\lambda\mu^{2i}}{\phi_0\phi_2 - \phi_1^2}(\phi_2 - 2i\phi_1 + i^2\phi_0).$$

The relative variance of the estimated frequency is therefore

$$\frac{\phi_2 - 2i\phi_1 + i^2\phi_0}{\lambda(\phi_0\phi_2 - \phi_1^2)}.$$

VII. TESTS OF DEPARTURE FROM PROPORTIONALITY

Before any use is made of data to which proportional frequencies have been fitted, it is necessary to test whether the observations depart significantly from the estimated frequencies.

Following Cochran [1954], we shall denote by X^2 the statistic used for testing the discrepancy between observed and expected frequencies, and by χ^2 the random variable with the familiar distribution, to which the distribution of X^2 approximates. We consider here the adequacy of the χ^2-approximation in testing X^2 for departure from expectation in the present problem.

Alternatively, we may use the fact that, because N and T are sufficient for λ and μ, the probability of a sample, conditional on N and T, is independent of the parameter values. An exact conditional test can thus be made, provided there are several possible samples corresponding to the given values of N and T. The probability of a sample of values $n_0, n_1, \cdots, n_{k-1}$ with $\sum n_i = N$ and $\sum in_i = T$ is

$$\frac{1}{\prod (n_i\,!)} e^{-\lambda\phi_0(\mu)}\lambda^N \mu^T.$$

Likewise the joint probability of N and T is

$$\frac{C(N, T)}{N!} e^{-\lambda\phi_0(\mu)}\lambda^N \mu^T,$$

where $C(N, T)$ is as defined in (2). Hence, the probability of the sample conditional on N and T is

$$\frac{N!}{C(N, T) \prod (n_i\,!)}. \tag{12}$$

By enumerating all possible samples and their probabilities we can decide the significance of an observed sample. Often the set of samples deemed to show significant departure from proportionality will be taken as a set of the samples with the smallest probabilities; however, in many problems, other criteria may be considered more appropriate. The use of X^2 has the advantage that it orders the samples according to a quantitative measure of departure from proportionality, whereas the probability attaching to a particular sample may be small merely because of the discontinuity of the distribution. A satisfactory arrangement would therefore be to order the samples by means of X^2, but to use the exact probabilities. In many cases the exact cumulated probabilities will not differ much from those given by the χ^2-distribution. The questions of calculation of probabilities and of choice of the significant set have been thoroughly discussed by Fisher [1950].

As an example of the exact calculation, we consider the following data in four classes:

	Observed	Expected
n_0	30	27
n_1	5	9
n_2	2	3
n_3	3	1

It will be seen that, because the numbers are restricted to four classes, the possibilities are much fewer in number than in other examples, such as that of the testing of deviations from a Poisson distribution, as discussed by Fisher. For this set, $N = 40$, $T = 18$, and it is readily verified that $\mu^* = \frac{1}{3}$, $\lambda^* = 27$. X^2 is easily calculated as

$$\frac{n_0^2}{27} + \frac{n_1^2}{9} + \frac{n_2^2}{3} + \frac{n_3^2}{1} - N = 6.444;$$

however, since the expected values in two of the classes are small, the tabulated distribution of χ^2 is not likely to be a good approximation to the distribution of X^2 for this sample. For this reason, as an example of method, the exact probabilities have been determined for the 37 possible samples having values $N = 40$, $T = 18$. The probabilities and the values of X^2 are given in Table 2.

We have $C(40, 18) = 220{,}495{,}831 \times 10^6$.

From the Table it is seen that the sample in question just attains significance at the 5 percent level; this is so whether the samples are ordered according to their probability or by X^2. The tabulated distribution would give a probability, for χ^2 with two degrees of freedom,

$$e^{-\frac{1}{2}X^2} = e^{-3.222} = 0.039866.$$

This shows that, for this example, the tabulated distribution underestimates the significance probability only slightly. On the other hand, for the sample

$$31 \quad 3 \quad 3 \quad 3$$

for which the probability accumulated according to X^2 is 0.008190,

$$X^2 = 8.593,$$

the probability of exceeding which is

$$0.013619.$$

The probability here is overestimated somewhat.

As a matter of interest, the mean and variance of for the distribution generated by this set are $E(X^2) = 1.9956$, $V(X^2) = 3.4223$, compared with the values of 2 and 4 respectively for the χ^2-distribution. The low variance will usually lead to overestimation of probabilities at the tails of the distribution and underestimation of significance, when referred to χ^2, though this effect will sometimes be masked by local irregularities of the distribution.

VIII. MOMENTS OF THE NON-PARAMETRIC DISTRIBUTION

We now show how the moments of the exact distribution of X^2

TABLE 2.

The Frequencies in Four Classes for which $N = 40$, $T = 18$

n_0	n_1	n_2	n_3	Probability	Cumulative probability	27 χ^2
34	0	0	6	.000000	.000000	1048
31	0	9	0	.000001	.000001	610
33	0	3	4	.000003	.000004	522
33	1	1	5	.000004	.000008	696
32	0	6	2	.000010	.000018	376
32	3	0	5	.000019	.000037	646
31	1	7	1	.000089	.000126	352
32	1	4	3	.000098	.000224	334
32	2	2	4	.000146	.000370	424
30	2	8	0	.000173	.000543	408
22	18	0	0	.000514	.001058	376
31	4	1	4	.000781	.001839	370
30	6	0	4	.000807	.002646	360
31	2	5	2	.000938	.003584	226
31	3	3	3	.002083	.005667	232
30	3	6	1	.003229	.008896	198
29	4	7	0	.003460	.012356	250
24	15	0	1	.004561	.016917	198
28	9	0	3	.005574	.022491	190
23	16	1	0	.006841	.029332	226
26	12	0	2	.009578	.038910	136
30	5	2	3	.009688	.048598	174
30	4	4	2	.012110	.060708	120
29	7	1	3	.013840	.074547	160
28	6	6	0	.023412	.097959	136
29	5	5	1	.029063	.127022	88
24	14	2	0	.034206	.161228	120
25	13	1	1	.038311	.199539	88
27	10	1	2	.046824	.246363	66
29	6	3	2	.048439	.294802	58
27	8	5	0	.070236	.365038	66
28	8	2	2	.075253	.440291	40
25	12	3	0	.083006	.523297	58
28	7	4	1	.100337	.623634	22
26	10	4	0	.105354	.728988	40
26	11	2	1	.114932	.843920	22
27	9	3	1	.156080	1.000000	0

may be determined. This distribution is independent of the population parameters, and depends only on the values of N and T.

We consider first the conditional moments of the n_i. From expression (12) we see that

$$\frac{C(N, T)}{N!} = \sum \frac{1}{\prod (n_i !)},$$

where the sum is taken over all sets of the n_i such that $\sum n_i = N$ and $\sum in_i = T$. Now the expected value of n_j is

$$\frac{N!}{C(N, T)} \sum \frac{n_j}{\prod (n_i !)}.$$

In the sum, the total of the arguments, with n_j replaced by $n_j - 1$, is $N - 1$, and the weighted sum of the arguments is $T - j$. It follows that $E(n_j) = NC(N - 1, T - j)/C(N, T)$. By a similar method it follows that, if $\sum a_i = A$, and $\sum ia_i = B$, then the general factorial moment is

$$E\left[\frac{\prod (n_i !)}{\prod (n_i - a_i)!}\right] = \frac{N!}{(N - A)!} \frac{C(N - A, T - B)}{C(N, T)}.$$

Thus in principle the joint factorial moments of the n_i can be determined provided we can regard the $C(N, T)$ as known. Some discussion of the asymptotic representation of $C(N, T)$ is given in Section X, but more needs to be known about these coefficients in general.

For the test of departure from expected frequencies we have

$$X^2 = \sum \frac{n_i^2}{\lambda^* \mu^{*i}} - N.$$

Now λ^* and μ^* are functions of N and T only, so are fixed for the conditional distribution. Hence X^2 may be simply regarded as a weighted sum of squares of the n_i. In particular, since

$$E(n_i^2) = \frac{N(N - 1)C(N - 2, T - 2i) + NC(N - 1, T - i)}{C(N, T)}$$

we have

$$E(X^2) = \frac{N}{\lambda^* C(N, T)}$$

$$\cdot \sum \frac{(N - 1)C(N - 2, T - 2i) + C(N - 1, T - i)}{\mu^{*i}} - N;$$

higher moments may be found similarly.

IX. TEST FOR TREND IN FREQUENCIES

Of particular importance is the test for the reality of any apparent trend in the frequencies—that is, whether the estimated value of μ differs significantly from unity. When N is large, the test may be made by comparing the difference $1 - \mu^*$ with its standard error, or, what is equivalent, testing $\frac{1}{2}(k - 1)N - T$ against its standard error.

When $\mu = 1$, the variance of μ^* reduces to $12/(k^2 - 1)N$, so that the test statistic, distributed approximately as χ^2 with 1 degree of freedom, is

$$(k^2 - 1)N(1 - \mu^*)^2/12. \tag{13}$$

More directly, the variance of T is

$$[(k^2 - 1)N]/12,$$

so that an alternative test statistic is

$$12(\tfrac{1}{2}(k - 1)N - T)^2/(k^2 - 1)N. \tag{14}$$

The statistics (13) and (14) are equivalent in large samples since, when μ^* is near to unity, we have

$$\mu^* = 1 - \frac{12(\frac{1}{2}(k - 1)N - T)}{(k^2 - 1)N} + O(N^{-1}),$$

as can be seen from (9).

The exact significance probability is given by the probability of a value of T less than or equal to that observed, which is easily computed directly when N is not large. This probability should be doubled, since both tails of the distribution are relevant to this test. With $\mu = 1$, this probability is simply

$$\sum_{r=0}^{T} C(N, r)/k^N = S(N, T)/k^N.$$

Now the generating function of $S(N, T)$ is

$$[\phi_0(\mu)]^N/(1 - \mu) = (1 - \mu^k)^N/(1 - \mu)^{N+1}.$$

From this representation we readily deduce that

$$S(N, T) = \binom{N + T}{N} - \binom{N}{1}\binom{N + T - k}{N} + \binom{N}{2}\binom{N + T - 2k}{N} - \cdots \tag{15}$$

which provides the most convenient means of computation of isolated values of the sum. On putting $T = (k - 1)N$ in (15) we have the interesting corollary

$$\binom{kN}{N} - \binom{N}{1}\binom{k(N - 1)}{N} + \binom{N}{2}\binom{k(N - 2)}{N} - \cdots = k^N.$$

Values of T less than $\frac{1}{2}(k - 1)N$ required for significance at the 5 and 1 percent levels of probability have been determined for various values of k and N, and are presented in Table 3. For values of N beyond the range of the Table, the large-sample χ^2-test may be used.

TABLE 3

EXACT TEST FOR EXISTENCE OF TREND
(I.E. OF NULL HYPOTHESIS $\mu = 1$)

5 PERCENT POINT OF T [OR $(k - 1)N - T$]

	k		
N	3	4	5
5	1	2	3
6	1	3	4
7	2	4	6
8	3	5	7
9	3	6	9
10	4	7	10
11	5	8	12
12	6	9	13
13	6	11	15
14	7	12	17
15	8	13	18

1 PERCENT POINT OF T [OR $(k - 1)N - T$]

	k		
N	3	4	5
5	0	0	1
6	0	1	2
7	1	2	4
8	1	3	5
9	2	4	6
10	3	5	8
11	3	6	9
12	4	7	11
13	5	8	12
14	5	9	14
15	6	11	15

X. EVALUATION OF TAIL PROBABILITIES

When N is large, the calculation of the probabilities of the tails of the distribution is troublesome. Blackwell and Hodges [1959] have given a method for expeditiously finding approximate tail probabilities. The method depends on a transformation of the distribution to a new distribution for which probabilities in the neighbourhood of the mean are to be determined. Daniels [1954] has given equivalent results, expressed explicitly for continuous probability densities, and Good [1957] has given similar results to the term of order N^{-2} of the leading term.

We briefly outline Blackwell and Hodges' results applied to the present distribution. We make use of the fact that the distribution of T is the N-fold convolution of the distribution of X as defined in Section IV.

The moment-generating function of $X - a$, where a is any constant, is $M(s) = E[e^{s(X-a)}]$.

Let s^* be the unique value of s that minimizes $M(s)$, and denote the minimum of $M(s)$ by $m(a)$; note that s^* is a function of a. Then we transform to a new variable Y, for which

$$P(Y = x) = \frac{P(X = x)e^{s^*(x-a)}}{m(a)}. \qquad (16)$$

It is readily seen that Y has in fact a probability density with the same range as X. Also, by differentiating the numerator of (16) with respect to s^* and summing, we find that $E(Y) = a$.

Before continuing with Blackwell and Hodges' method, we establish the interesting result that, for any distribution for which the sum of the values of X is a sufficient statistic for the parameter μ, the new variate Y has a distribution of the same form, but with parameter μ^*, the maximum likelihood estimate corresponding to the value $X = a$. This result is probably well known, though we have never seen it discussed. Thus we see that, for distributions of this commonly occurring class, the solution of the maximum likelihood equation is important not only for providing a point estimate, but also for evaluating the tail probabilities corresponding to the particular observed value.

If the sum of the values of X is sufficient for μ, the density of X may be written as

$$h(X)e^{Xg(\mu)}/f(\mu), \qquad (17)$$

where

$$f(\mu) = \sum_x h(x)e^{xg(\mu)}.$$

Then

$$M(s) = \sum_x h(x)e^{s(x-a)+xg(\mu)}/f(\mu),$$

so that, for any value of s whatever, a transformed density is given by

$$h(X)e^{X[s+g(\mu)]}/e^{as}M(s).$$

This density is clearly of the same form as that of X.

Now the maximum likelihood equation, corresponding to an observed value a, turns out to be

$$a - \sum_x xh(x)e^{xg(\mu)}/\sum_x h(x)e^{xg(\mu)} = 0,$$

where we have differentiated (17) with respect to $g(\mu)$ rather than μ itself. This is in fact $E(X \mid \mu = \mu^*) = a$. Now from (16) we derived that $E(Y) = a$. It follows therefore that the parameter of the distribution of Y is μ^*.

We can also express $m(a)$ in terms of the maximum likelihood estimator. On putting $x = a$ in (16), we see that $m(a) = P(X = a)/P(Y = a)$, or $P(a \mid \mu)/P(a \mid \mu^*)$.

For the particular distribution we have been considering, therefore

$$m(a) = \left(\frac{\mu}{\mu^*}\right)^a \left(\frac{\phi_0^*}{\phi_0}\right),$$

where we indicate by an asterisk a function of μ^*.

Blackwell and Hodges, by means of the transformation (16), and taking $Na = T$, show that

$$P(X_1 + X_2 + \cdots + X_N = T) = [m(a)]^N P(Y_1 + Y_2 + \cdots + Y_N = T),$$

and thence, that each side is equal to

$$\frac{[m(a)]^N}{\sqrt{2\pi N \kappa_2^*}} \left[1 + \frac{1}{8N}\left(\frac{\kappa_4^*}{\kappa_2^{*2}} - \frac{5}{3}\frac{\kappa_3^{*2}}{\kappa_2^{*2}}\right) + O(N^{-2})\right],$$

or

$$\left(\frac{\mu}{\mu^*}\right)^T \left(\frac{\phi_0^*}{\phi_0}\right)^N \frac{1}{\sqrt{2\pi N \kappa_2^*}} \left[1 + \frac{1}{8N}\left(\frac{\kappa_4^*}{\kappa_2^{*2}} - \frac{5}{3}\frac{\kappa_3^{*2}}{\kappa_2^{*3}}\right) + O(N^{-2})\right] \qquad (18)$$

where the κ_r^* are the cumulants of the distribution of Y; that is, functions of μ^*. The explicit terms in (18) give a satisfactory approximation to $P(T \mid N)$, provided N is large.

We note that the series in (18) depends only on μ^* (i.e., on T), and not on μ. Thus what we have obtained is an asymptotic expansion of $C(N, T)$. In fact,

$$C(N,\,T) = \frac{\phi_0^{*N}}{\mu^{*T}\sqrt{2\pi N\,\kappa_2^*}}\left[1 + \frac{1}{8N}\left(\frac{\kappa_4^*}{\kappa_2^{*2}} - \frac{5}{3}\,\frac{\kappa_3^{*2}}{\kappa_2^{*3}}\right) + O(N^{-2})\right].$$

Thus, $C(N,\,T)$ may be estimated by substituting the actual values of the cumulants, from the formulae given in Section IV.

As a check on the asymptotic formula, we estimate $C(N,\,T)$ for the values $k = 4$, $N = 15$, $T = 11$. We deduce $\mu^* = \frac{1}{2}$, $\phi_0 = 15/8$,

$$\kappa_2^* = 194/15^2 = 0.862222,$$

$$\kappa_3^* = 2842/15^3 = 0.842074,$$

$$\kappa_4^* = 1434/15^4 = 0.028326.$$

Hence

$$C(15,\,11) \sim \sqrt{\frac{15}{2\pi\cdot 194}}\,\frac{(15/8)^{15}}{(\frac{1}{2})^{11}}\left(1 - \frac{1.805604}{120}\right) = 2{,}784{,}963.$$

The true value is

$$\binom{25}{15} - \binom{15}{1}\binom{21}{15} + \binom{15}{2}\binom{17}{15} = 2{,}784{,}600,$$

from which the estimate by the asymptotic formula is in error by 130 per million.

For significance testing, the cumulative tail probabilities are more important than the individual terms. These cumulative probabilities are also given by Blackwell and Hodges. They give

$$\begin{aligned}P(T \geq t) &= \frac{[m(a)]^N}{\sqrt{2\pi N\,\kappa_2^*}(1 - z)}\left[1 + \frac{1}{8N}\left(\frac{\kappa_4^*}{\kappa_2^{*2}} - \frac{5}{3}\,\frac{\kappa_3^{*2}}{\kappa_2^{*3}}\right)\right.\\ &\qquad\left. - \frac{z}{2N}\left(\frac{\kappa_3^*(1 - z) + \kappa_2^*(1 + z)}{\kappa_2^{*2}(1 - z)^2}\right) + O(N^{-2})\right]\\ &= \frac{P(T = t)}{(1 - z)}\left[1 - \frac{z}{2N}\,\frac{\kappa_3^*(1 - z) + \kappa_2^*(1 + z)}{\kappa_2^{*2}(1 - z)^2} + O(N^{-2})\right]\end{aligned}$$

where, for the distribution considered here, $z = \mu/\mu^*$. We note that, as is to be expected, the bracketed series in the cumulative probability depends on μ. The approximation is valid only for $\mu^* > \mu$, and is most effective when z is small—that is, when μ^* is large and the extreme tails of the distribution are being considered.

When we are determining the probability in the lower tail of the distribution, that is $P(T \leq t)$, so that $T < \frac{1}{2}(k - 1)N$, and $\mu^* < \mu$, we simply replace z by z^{-1} throughout the formula, and change the sign of κ_3^*.

As a test of the adequacy of this approximation, we apply it to the example for which the individual term was calculated above, namely $k = 4$, $N = 15$, $T = 11$.

Inserting the values found above, we have $z = 2\mu$,

$$P(T \leq 11) = \frac{P(T = 11)}{\left(1 - \dfrac{1}{2\mu}\right)}$$

$$\cdot \left[1 - \frac{1}{60\mu} \left\{ \frac{\dfrac{-2842}{15^3}\left(1 - \dfrac{1}{2\mu}\right) + \dfrac{194}{15^2}\left(1 + \dfrac{1}{2\mu}\right)}{\dfrac{194^2}{15^4}\left(1 - \dfrac{1}{2\mu}\right)^2} \right\} + O(N^{-2}) \right].$$

We consider the adequacy of the approximation when $\mu = 1$, for which we can readily determine the exact probability. We have

$$P(T = 11) = 2{,}784{,}600/4^{15},$$

so the leading term in the approximate probability is

$$2P(T = 11) = 5{,}569{,}200/4^{15} = 0.005187.$$

For the first adjustment we find the factor $1 - 0.078223$ giving $5{,}133{,}560/4^{15} = 0.004781$. The exact value of the denominator is

$$S(15, 11) = \binom{26}{15} - \binom{15}{1}\binom{22}{15} + \binom{15}{2}\binom{18}{15} = 5{,}253{,}680,$$

giving the exact probability 0.004893.

We see that the first adjustment gives an approximation adequate for the purpose of assessing the significance probability. Since both tails of the distribution are relevant to any test of significance, the significance probability is 0.009786; $T = 11$ is thus the 1 percent point of the distribution.

By determining the tail probabilities corresponding to any given value of μ we can in principle set a confidence range for μ — the set of all μ which are not discordant with the observed values N and T.

XI. APPLICATION TO EXPERIMENTAL DATA

The data in Table 4 show numbers in each of the first three instars of immature cowpea aphids (*A. craccivora*), from samples drawn on six different occasions from the same population. Since the experimental conditions were uniform, and the average duration of each instar is the same (about 42 hours at 20°C), the expected numbers are in geometric progression. The Table also gives an overall estimate of the common ratio, 0.529820. In order to determine the growth-rate of the population,

TABLE 4

AGE-DISTRIBUTION OF IMMATURE COWPEA APHIDS (*A. craccivora*)

	Instar							
Sample	I	II	III	N	T	λ^*	μ^*	X^2
1	181	92	51	324	194	179.730	0.526015	0.11
2	148	78	42	268	162	147.737	0.531518	0.01
3	130	54	43	227	140	123.456	0.543412	4.07
4	70	37	21	128	79	69.580	0.543848	0.03
5	88	41	20	149	81	87.710	0.474049	0.13
6	85	42	28	155	98	82.857	0.558629	0.63
Total	702	344	205	1251	754	690.958	0.529820	2.14

we equate this ratio to $e^{-c\rho}$, c being the average duration of instar and ρ the growth-rate.

We then find

$$c\rho = -\log(0.529820) = 0.6352$$

so that

$$\rho = 0.6352/42 \text{ per hour} = 0.01512 \text{ per hour} = 0.363 \text{ per day}.$$

The population grows at the rate of $100(e^{0.363} - 1) = 44$ percent per day.

Table 4 also gives estimates of λ and μ for each sample, and X^2 with 1 degree of freedom measuring departure from the common ratio. Only the largest X^2-value, for sample 3, attains the 5 percent level of significance, so that there is no evidence for departure from hypothesis.

The variance of the estimate of μ from a sample of N is $\mu^2/V(T)$ which reduces when $k = 3$ to

$$\frac{\mu(1 + \mu + \mu^2)^2}{N(1 + 4\mu + \mu^2)}.$$

Using the overall estimate $\mu_0^* = 0.529820$ we find the variance of estimate to be

$$V(\mu^* \mid N) = 0.510813/N.$$

The consistency of the estimates of μ from the different samples may be tested by X^2.

For consistency,

$$X^2 = (0.526015^2 \times 324 + 0.531518^2 \times 268 + \cdots - 0.529820^2 \times 1251)/0.510813$$
$$= 1.84, \text{ with 5 degrees of freedom.}$$

The individual ratios differ no more than would be expected by chance.

After the manner of Section IX, formulae (13) and (14), an alternative statistic for testing consistency is based on comparison of the ratios T/N.

$$V(T/N) = \mu(1 + 4\mu + \mu^2)/N(1 + \mu + \mu^2)^2 = 0.549534/N$$

Hence

$$X^2 = (194^2/324 + 162^2/268 + \cdots - 754^2/1251)/0.549534$$
$$= 1.33.$$

REFERENCES

Blackwell, David and J. L. Hodges, Jr. [1959]. The probability in the extreme tail of a convolution. *Ann. Math. Stat. 30*, 1113–20.

Chapman, D. G., and D. S. Robson [1960]. The analysis of a catch curve. *Biometrics 16*, 354–68.

Cochran, W. G. [1954]. Some methods for strengthening the common χ^2 tests. *Biometrics 10*, 417–51.

Daniels, H. E. [1954]. Saddlepoint approximations in statistics. *Ann. Math. Stat. 25*, 631–50.

Fisher, R. A. [1950]. The significance of deviations from expectation in a Poisson series. *Biometrics 6*, 17–24.

Good, I. J. [1957]. Saddle-point methods for the multinomial distribution. *Ann. Math. Stat. 28*, 861–81.

Noack, Albert [1950]. A class of random variables with discrete distributions. *Ann. Math. Stat. 21*, 127–32.

Patil, G. P. [1961]. On homogeneity and combined estimation for generalized power series distribution and certain applications. *Biometrics 18,* to be published.

Riordan, John [1958]. *An Introduction to Combinatorial Analysis.* New York; Wiley.

COMPUTING PROCEDURES FOR ESTIMATING COMPONENTS OF VARIANCE IN THE TWO-WAY CLASSIFICATION, MIXED MODEL[1]

S. R. Searle,
New Zealand Dairy Board, Wellington, New Zealand,
AND
C. R. Henderson
Cornell University, Ithaca, N. Y., U. S. A.

INTRODUCTION

Methods for estimating variance components from unbalanced data are given in Henderson [1953] for both the random model and the mixed model. The calculations involved in the latter case are somewhat tedious and usually not computationally feasible for data having many classes. This paper outlines the analysis for unbalanced data from a two-way classification with one classification considered fixed. Simplified computing procedures are presented suitable for a large number of levels in the random classification and a reasonable number of levels of the fixed classification.

MODELS AND ESTIMATION

The model for the two-way classification can be taken as

$$x_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk} \ , \qquad (1)$$

where x_{ijk} is the observation and μ is the general mean. α_i is the effect due to the i'th level of the α-class, β_j the effect due to the j'th level of the β-class, $\alpha\beta_{ij}$ the interaction and e_{ijk} a random error term. We will suppose that the number of α-classes in the data is a, the number of β-classes is b, that there are n_{ij} observations in the ij'th subclass, and that s sub-classes have observations in them. All terms except μ are taken as independent random variables in the random model with zero means and variances σ_α^2 , σ_β^2 , $\sigma_{\alpha\beta}^2$ and σ_e^2 respectively. The β-classification will be considered fixed in the mixed model, the variances involved being σ_α^2 , $\sigma_{\alpha\beta}^2$, and σ_e^2 , the interaction terms being random for the particular set of fixed effects occurring in the data. In this case we are assuming an underlying multivariate distribution with x_{ijk} having mean

[1] Parts of this paper are based on the thesis submitted by the first author in partial fulfillment of the requirements of the Ph. D. degree.

value $\mu + \beta_j$ and variance $\sigma_\alpha^2 + \sigma_{\alpha\beta}^2 + \sigma_e^2$: the covariance between x_{ijk} and $x_{ijk'}$ for $k \neq k'$ is $\sigma_\alpha^2 + \sigma_{\alpha\beta}^2$ and that between x_{ijk} and $x_{ij'k'}$ for $j \neq j$ and $k \neq k'$, is σ_α^2 .

Variance components can be estimated in both models by equating linear functions of sums of squares to their expected values, the sums of squares used being reductions in the total sum of squares due to fitting various elements of the model as if it were a fixed model. For example, fitting $(\mu + \alpha_i)$ results in matrix equations of the form

$$\mathbf{P}\hat{\boldsymbol{\alpha}} = \mathbf{y},$$

where $\mathbf{y}$ is the vector of α-class totals $x_{i..}$, and $\hat{\boldsymbol{\alpha}}$ is the vector of estimates of the α's. $\mathbf{P}$ is non-singular, a diagonal matrix of order a, of terms $n_{i.}$. Thus

$$\hat{\boldsymbol{\alpha}} = \mathbf{P}^{-1}\mathbf{y},$$

and the reduction in the total sum of squares due to fitting the α's is

$$R(\mu, \alpha) = \hat{\boldsymbol{\alpha}}\mathbf{y} = \mathbf{y}'\mathbf{P}^{-1}\mathbf{y}.$$

This is the usual uncorrected sum of squares, namely

$$R(\mu, \alpha) = \sum_i x_{i..}^2/n_{i.} \ . \tag{2}$$

Similarly

$$R(\mu, \beta) = \sum_j x_{.j.}^2/n_{.j} \ , \tag{3}$$

$$R(\mu, \alpha, \beta, \alpha\beta) = \sum_i \sum_j x_{ij.}^2/n_{ij} \ , \tag{4}$$

$$R(\mu) = x_{...}^2/n_{..} \ , \tag{5}$$

and

$$R(0) = \sum_i \sum_j \sum_k x_{ijk}^2 \ . \tag{6}$$

The variance components of the random model are estimated by equating the expected values to the observed values of differences among the above uncorrected sums of squares, namely (2)–(5), (3)–(5), (4)–(2)–(3) + (5) and (6)–(4). These cannot be used in the mixed model because, apart from (6)–(4), their expectations then contain functions of the fixed effects. The within-subclasses sum of squares, (6)–(4), can still be used to estimate the error variance as in the random model, and σ_α^2 and $\sigma_{\alpha\beta}^2$ can be estimated from the differences:

$$R(\mu, \alpha, \beta) - R(\mu, \beta),$$

and

$$R(\mu, \alpha, \beta, \alpha\beta) - R(\mu, \alpha, \beta),$$

whose expectations are free of terms in the fixed effects. $R(\mu, \alpha, \beta)$ is the reduction in the total sum of squares due to fitting α and β alone without the interaction terms. This reduction in sum of squares does not occur in the random model analysis and, when required for the mixed model analysis it usually involves a considerable amount of computing. Its relationship to the random model analysis will now be shown and a simplified computing procedure derived.

SIMPLIFIED EXPRESSION FOR R (μ, α, β)

Henderson *et al.* [1959] have shown that fixed effects in a mixed model can be estimated by treating the random effects as fixed and maximizing the joint distribution function of the observations and the random effects. The resulting equations for the two-way classification without interaction are

$$\begin{bmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{R} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \tag{7}$$

where P is a diagonal matrix, of order a, with terms $n_{i.}$, $\mathbf{R}$ is a diagonal matrix of order $b - 1$ with terms $n_{.j}$, $j \neq b$, $\mathbf{Q}$ is a matrix of order a by $b - 1$ with terms n_{ij} , $j \neq b$, $\mathbf{y}$ is a vector of the $x_{i..}$ totals and $\mathbf{z}$ is a vector of the $x_{.j.}$ totals. $\mathbf{R}$, $\mathbf{Q}$ and $\mathbf{z}'$ have $b - 1$ columns because of omitting the equation for the last β, appropriate to imposing the constraint $\hat{\beta}_b = 0$ in order to have a unique solution. $R(\mu, \alpha, \beta)$ from these equations is

$$\begin{aligned} R(\mu, \alpha, \beta) &= (\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\beta}}') \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}, \\ &= (\mathbf{y}'\mathbf{z}') \begin{bmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{R} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}. \end{aligned} \tag{8}$$

This expression for obtaining $R(\mu, \alpha, \beta)$ requires inverting a matrix of order $(a + b - 1)$, which is not feasible in many situations because a, the number of random classes in the data, is large. For example, the analysis of dairy production records in Searle and Henderson [1960] involved 688 herds (random) and 4 age groupings (fixed) so that a matrix of order 691 would need to be inverted for equation (8). However, because of its special form, it can be reduced to inverting a 3 × 3 matrix. Thus in the general case for the two-way classification the inversion of an $(a + b - 1)$ matrix can be reduced to inverting one of order $(b - 1)$ and this is a considerable reduction in the computing required especially when a, the number of random effects, is large, as is frequently the case.

The simplification of (8) proceeds as follows. Let

$$\begin{bmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{R} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{bmatrix}. \tag{9}$$

Then

$$\begin{aligned} \mathbf{A} &= \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{Q}\mathbf{D}\mathbf{Q}'\mathbf{P}^{-1}, \\ \mathbf{B} &= -\mathbf{P}^{-1}\mathbf{Q}\mathbf{D}, \\ \mathbf{D} &= (\mathbf{R} - \mathbf{Q}'\mathbf{P}^{-1}\mathbf{Q})^{-1}, \end{aligned} \tag{10}$$

$$\begin{aligned} R(\mu, \alpha, \beta) &= (\mathbf{y}'\mathbf{z}')\begin{bmatrix} \mathbf{P}^{-1} + \mathbf{P}^{-1}\mathbf{Q}\mathbf{D}\mathbf{Q}'\mathbf{P}^{-1} & -\mathbf{P}^{-1}\mathbf{Q}\mathbf{D} \\ -\mathbf{D}\mathbf{Q}'\mathbf{P}^{-1} & \mathbf{D} \end{bmatrix}\begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \\ &= \mathbf{y}'\mathbf{P}^{-1}\mathbf{y} + (\mathbf{z}' - \mathbf{y}'\mathbf{P}^{-1}\mathbf{Q})\mathbf{D}(\mathbf{z} - \mathbf{P}^{-1}\mathbf{Q}'\mathbf{y}). \end{aligned} \tag{11}$$

The matrix $\mathbf{P}$ is easily inverted, being diagonal, and the only non-diagonal matrix to invert is $\mathbf{D} = (\mathbf{R} - \mathbf{Q}'\mathbf{P}^{-1}\mathbf{Q})^{-1}$ of order $b - 1$.

The first term in the above expression is $R(\mu, \alpha)$, and the second can be derived from equations (7). Eliminating

$$\hat{\boldsymbol{\alpha}} = \mathbf{P}^{-1}(\mathbf{y} - \mathbf{Q}\hat{\boldsymbol{\beta}}),$$

gives

$$\hat{\boldsymbol{\beta}} = \mathbf{D}(\mathbf{z} - \mathbf{Q}'\mathbf{P}^{-1}\mathbf{y}). \tag{12}$$

This is easily computed and can then be used to obtain R_β, the reduction in the sum of squares due to fitting the β's in this no-interaction model, as

$$R_\beta = \hat{\boldsymbol{\beta}}'(\mathbf{z} - \mathbf{Q}'\mathbf{P}^{-1}\mathbf{y}). \tag{13}$$

Thus from (11) $R(\mu, \alpha, \beta)$ can be expressed as

$$R(\mu, \alpha, \beta) = R(\mu, \alpha) + R_\beta , \tag{14}$$

the first term of which is part of the random model analysis and the second term comes from (12) and (13).

EXPECTED VALUES

The expectation of $R(\mu, \alpha, \beta)$ can be found by using (8) and (9) and writing

$$\begin{aligned} R(\mu, \alpha, \beta) &= (\mathbf{y}'\mathbf{z}')\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{D} \end{bmatrix}\begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}, \\ &= (\mathbf{y}'\mathbf{A}\mathbf{y} + \mathbf{z}'\mathbf{B}\mathbf{y} + \mathbf{y}'\mathbf{B}\mathbf{z} + \mathbf{z}'\mathbf{D}\mathbf{z}), \\ &= \operatorname{tr}(\mathbf{A}\mathbf{y}\mathbf{y}' + 2\mathbf{B}\mathbf{z}\mathbf{y}' + \mathbf{D}\mathbf{z}\mathbf{z}'), \end{aligned}$$

where $tr(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$, the sum of its diagonal terms. After substituting from (1) for the vectors of totals $\mathbf{y}$ and $\mathbf{z}$ and also using (1) in (3), (4) and (6) it can be shown, with a certain amount of algebraic manipulation, that the differences used for estimating the variance components have expected values

$$E[R(\mu, \alpha, \beta) - R(\mu, \beta)]$$
$$= (n_{..} - k_2)\sigma_\alpha^2 + (k_\beta - k_2)\sigma_{\alpha\beta}^2 + (a - 1)\sigma_e^2 ,$$

$$E[R(\mu, \alpha, \beta, \alpha\beta) - R(\mu, \alpha, \beta)]$$
$$= \qquad (n_{..} - k_\beta)\sigma_{\alpha\beta}^2 + (s - a - b + 1)\sigma_e^2 ,$$

$$E[R(0) - R(\mu, \alpha, \beta, \alpha\beta)]$$
$$= \qquad (n_{..} - s)\sigma_e^2 .$$

The variances are estimated by equating the right-hand sides of these equations to the calculated values of the differences in the square brackets of the left-hand side. The term k_2 is one of the coefficients used in the random model analysis, namely $k_2 = \sum_{j=1}^{b} [\sum_{i=1}^{a} n_{ij}^2]/n_{.j}$. The coefficient k_β comes from the expected value of $R(\mu, \alpha, \beta)$ and can be expressed in the form

$$k_\beta = 2 \text{ tr } (\mathbf{AU} + 2\mathbf{BV} + \mathbf{DW}), \tag{15}$$

where $\mathbf{U}$ is a diagonal matrix of order a with terms $\sum_{j=1}^{b-1} n_{ij}^2$ for $i = 1 \cdots a$, $\mathbf{V}$ is a matrix of order a by $b - 1$ with terms n_{ij}^2 for $i = 1 \cdots a$, and $j = 1 \cdots b - 1$, and W is a diagonal matrix of order $b - 1$ with terms $\sum_{i=1}^{a} n_{ij}^2$ for $j = 1 \cdots b - 1$.

COMPUTING R_β AND k_β

R_β is obtained from computing the following terms:

(i) A square matrix $\mathbf{C}$, of order $(b - 1)$ with terms

$$c_{jj} = n_{.j} - \sum_i n_{ij}^2/n_{i.} , \qquad j = 1, \cdots , b - 1,$$

and

$$c_{jj'} = -\sum_i n_{ij}n_{ij'}/n_{i.}, \qquad j \neq j' = 1, \cdots , b - 1.$$

Computing c_{bj} provides the check that $\sum_{j'=1}^{b} c_{jj'} = 0$, for all j, i.e. row and column totals of the augmented $\mathbf{C}$ are zero.

(ii) A column vector $\mathbf{r}$, of order $b - 1$, whose terms are

$$r_j = x_{.j.} - \sum_i n_{ij}\bar{x}_{i..} , \qquad j = 1, \cdots , b - 1.$$

Computing r_b provides the check that the sum of all the r_j's is zero.

(iii) A column vector, of order $b - 1$, of the estimates of the β's,

$$\hat{\boldsymbol{\beta}} = \mathbf{C}^{-1}\mathbf{r}.$$

(iv) The "inner product" of the terms in the preceding two vectors is R_β:

$$R_\beta = \hat{\boldsymbol{\beta}}'\mathbf{r} = \sum_{j=1}^{b-1} \hat{\boldsymbol{\beta}}_j r_j = \mathbf{r}'\mathbf{c}\mathbf{r}.$$

This is equation (13), the $\mathbf{C}^{-1}$ used here being identical to $\mathbf{D}$.

Expression (15) for k_β is obtained using the elements of $\mathbf{C}^{-1}$, which we shall call $d_{jj'}$, obtained in the calculating of R_β above. The special forms of the matrices involved in (15) enables k_β to be written as

$$k_\beta = \sum_{i=1}^{a} \lambda_i + \sum_{j=1}^{b-1} d_{jj}\left(\sum_{i=1}^{a} f_{i,jj}\right) + 2 \sum_{j' \neq j, =1}^{b-1} \sum^{b-1} d_{jj'}\left(\sum_{i=1}^{a} f_{i,jj'}\right),$$

where

$$\lambda_i = \sum_{j=1}^{b-1} n_{ij}^2/n_{i.}\,,$$

$$f_{i,jj} = (n_{ij}^2/n_{i.})\,(\lambda_i + n_{i.} - 2n_{ij}),$$

and

$$f_{i,jj'} = (n_{ij}n_{ij'}/n_{i.})\,(\lambda_i - n_{ij} - n_{ij'}).$$

The f's, of which $\frac{1}{2}b(b - 1)$ in number are required, can be calculated for each i and summed, and the expression is well-suited to evaluation on a computer. Desk calculation can be arranged from the same formula when b is small, and the number of α-classes (random) is not too large. Calculating all of the $\frac{1}{2}b(b + 1)$ f's enables the following check to be placed on them:

$$\sum_{j'=1}^{b} \left(\sum_i f_{i,jj'}\right) = -\sum_i n_{ib}^2 n_{ij}/n_{i.}\,. \tag{16}$$

The f-values and $\sum_i \lambda_i$ can be obtained simultaneously with the terms of $\mathbf{C}$ and $\mathbf{r}$; R_β is calculated as $\mathbf{r}'\mathbf{C}^{-1}\mathbf{r}$ and k_β is obtained using the elements of $\mathbf{C}^{-1}$ as above.

EXAMPLE

The calculation of R_β and k_β is demonstrated for the small hypothetical example shown in Table 1.

The steps for obtaining R_β are as follows:

(i) The c-values are

$$c_{11} = 53 - 1^2/10 - 3^2/20 - 12^2/40 - 12^2/50 - 25^2/50 = 33.47,$$

$$c_{22} = 45 - 2^2/10 - 6^2/20 \qquad\qquad - 12^2/50 - 25^2/50 = 27.42,$$

TABLE 1
HYPOTHETICAL EXAMPLE

i	Number of observations n_{ij}					Mean observed values, $x_{i..}$
	j				Totals $n_{i.}$	
	1	2	3	4		
1	1	2	3	4	10	200
2	3	6	4	7	20	300
3	12	—	12	16	40	400
4	12	12	13	13	50	500
5	25	25	—	—	50	600
Totals $n_{.j}$	53	45	32	40	$n_{..} = 170$	
Totals of observed values, $x_{.j.}$	23000	23000	14000	19000		

$$c_{33} = 32 - 3^2/10 - 4^2/20 - 12^2/40 - 13^2/50 = 23.32,$$

$$c_{44} = 40 - 4^2/10 - 7^2/20 - 16^2/40 - 13^2/50 = 26.17,$$

$$c_{12} = -1(2)/10 - 3(6)/20 - 12(12)/50 - 25(25)/50 = -16.48,$$

$$c_{13} = -1(3)/10 - 3(4)/20 - 12(12)/40 - 12(13)/50 = -\ 7.62,$$

and similarly

$$c_{14} = -9.37, \qquad c_{24} = -\ 6.02,$$

$$c_{23} = -4.92, \qquad c_{34} = -10.78.$$

The check on these values, namely $\sum_{j'=1}^{b} c_{jj'} = 0$, can be seen to hold true; for example

$$c_{11} + c_{12} + c_{13} + c_{14} = 33.47 - 16.48 - 7.62 - 9.37 = 0.$$

(ii) The r-values are obtained using the totals of the observations for the levels of the fixed effects and the means for the levels of the random effects:

$$r_1 = 23000 - 1(200) - 3(300) - 12(400) - 12(500) - 25(600) = -3900,$$

$$r_2 = 23000 - 2(200) - 6(300) \qquad - 12(500) - 25(600) = -200,$$

$$r_3 = 14000 - 3(200) - 4(300) - 12(400) - 13(500) = 900,$$

$$r_4 = 19000 - 4(200) - 7(300) - 16(400) - 13(500) = 3200.$$

A check on these is that their sum is zero.

(iii)

$$\mathbf{C}^{-1} = \mathbf{D} = \begin{bmatrix} 33.47 & -16.48 & -7.62 \\ -16.48 & 27.42 & -4.92 \\ -7.62 & -4.92 & 23.32 \end{bmatrix}^{-1}$$

(iv)

$$R_\beta = \mathbf{r}'\mathbf{D}\mathbf{r}$$

$$= (-3900, -200, 900)\begin{bmatrix} .053824 & .036902 & .025373 \\ .036902 & .063205 & .025393 \\ .025373 & .025393 & .056530 \end{bmatrix}\begin{bmatrix} -3900 \\ -200 \\ 900 \end{bmatrix}$$

$$= 736703.$$

Calculating k_β involves the elements of **D** and the f-values. The latter, together with the λ_i-terms are shown in Table 2. The calculations for $i = 1$ are as follows.

$$\lambda_1 = (1^2 + 2^2 + 3^2)/10 = 1.40,$$

$$f_{1,11} = 1^2(1.4 + 10 - 2)/10 = .94,$$

$$f_{1,22} = 2^2(1.4 + 10 - 4)/10 = 2.96,$$

$$f_{1,33} = 3^2(1.4 + 10 - 6)/10 = 4.86,$$

$$f_{1,12} = 1(2)(1.4 - 1 - 2)/10 = -.32,$$

$$f_{1,13} = 1(3)(1.4 - 1 - 3)/10 = -.78.$$

Similarly $f_{1,14} = -1.44$, $f_{1,23} = -2.16$, $f_{1,24} = -3.68$ and $f_{1,34} = -6.72$. The sums of these terms, over all values of i, are shown in Table 2 from which the f-values can be checked by the expression (16); for example

$$\sum_i (f_{i,11} + f_{i,12} + f_{i,13} + f_{i,14})$$

$$= 505.8357 - 360.9718 - 113.1132 - 158.0607$$

$$= -126.31$$

$$= -4^2(1)/10 + 7^2(3)/20 + 16^2(12)/40 + 13^2(12)/50$$

$$= -(1.60 + 7.35 + 76.80 + 40.56).$$

From the c's and the f's k_β is now computed as

$$k_\beta = 45.79 + [505.8357(.053824) + 436.5532(.063205) + 212.4332(.056530)]$$

TABLE 2

Terms Used in Calculating k_β

i	λ_i	$f_{i,11}$	$f_{i,22}$	$f_{i,33}$	$f_{i,44}$	
1	1.40	.9400	2.9600	4.8600	5.4400	
2	3.05	7.6725	19.8900	12.0400	22.1725	
3	7.20	83.5200		83.5200	97.2800	
4	9.14	101.2032	101.2032	112.0132	112.0132	
5	25.00	312.5000	312.5000	—	—	
Total	45.79	505.8357	436.5532	212.4332	236.9057	
i	$-f_{i,12}$	$-f_{i,13}$	$-f_{i,14}$	$-f_{i,23}$	$-f_{i,24}$	$-f_{i,34}$
1	.3200	.7800	1.4400	2.1600	3.6800	6.7200
2	5.3550	2.3700	7.2975	8.3400	20.8950	11.1300
3	—	60.4800	99.8400	—	—	99.8400
4	42.7968	49.4832	49.4832	49.4832	49.4832	56.9868
5	312.5000	—	—	—	—	—
Total	360.9718	113.1132	158.0607	59.9832	74.0582	174.6768

$$\begin{aligned} &- 2[360.9718(.036902) + 113.1132(.025373) \\ &\qquad\qquad + 59.9832(.025393)] \\ &= 77.16. \end{aligned}$$

This procedure for obtaining R_β and k_β may not appear greatly more straight-forward than inverting the matrix required in (8) which in this case is the 8 × 8 matrix

$$\begin{bmatrix} 10 & & & & & 1 & 2 & 3 \\ & 20 & & & & 3 & 6 & 4 \\ & & 40 & & & 12 & 0 & 12 \\ & & & 50 & & 12 & 12 & 13 \\ & & & & 50 & 25 & 25 & 0 \\ 1 & 3 & 12 & 12 & 25 & 53 & & \\ 2 & 6 & 0 & 12 & 25 & & 45 & \\ 3 & 4 & 12 & 13 & 0 & & & 32 \end{bmatrix}.$$

Advantages are apparent however, when one considers the case of a large number of levels of the random classification 500 say, instead of 5. The matrix to be inverted would then be of order 503 while the procedure outlined here would still only require inverting a 3 × 3. The calculation of its terms the c's and of the terms for k_β, the f's, is still lengthy but can be accomplished separately for each i and summation made over i. This can be arranged quite straightforwardly for a desk calculator and is easily organized for an electronic computer such as the IBM 650, for example.

REFERENCES

Eisenhart, C. [1947]. The assumptions underlying the analysis of variance. *Biometrics 3*, 1–21.

Henderson C. R. [1953]. Estimation of variance and covariance components. *Biometrics 9*, 226–52.

Henderson C. R., Kempthorne O., Searle S. R., and Von Krosigk C. M. [1959]. Estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Searle S. R. [1958]. Sampling variances of estimates of components of variance. *Ann. Math. Stat. 29*, 167–78.

Searle S. R. and Henderson C. R. [1960]. Judging the effectiveness of age correction factors. *J. Dairy Sci. 43*, 966–74.

OPTIMUM SAMPLE SIZE IN ANIMAL DISEASE CONTROL[1]

A. W. Nordskog[2], H. T. David[3] and H. B. Eisenberg[4]
Iowa State University, Ames, Iowa, U. S. A.

1. INTRODUCTION

The objective of an animal disease control program would either be (*a*) to completely eradicate a disease or alternatively (*b*) to reduce the incidence of a disease to a low level and then to keep it in check. In the case of (*b*) the causative organism might not be completely stamped out.

If a disease has not obtained a strong foothold in a country, then objective (*a*) clearly is justified and perhaps the most drastic methods of control are in order. If, on the otherhand, a disease is widely distributed both geographically and zoologically, then alternative (*b*) might be more realistic. This implies that the objective of a particular control program should take into account the cost of the program relative to the price of the consequence if there were no control.

A case in point is the method of control for Pullorum disease (Salmonella pullorum) in poultry. The active form of the disease causes an enteritis in baby chicks which may result in high death losses. Chicks which recover may be carriers of the disease as adults. Female carriers may then transmit the organism via the egg to their progeny, where the disease again may become epidemic. Other species of domestic birds, and numerous wild species including pheasants, quail, and even foxes, cats, swine, cattle and man, have been reported as potential carriers of Pullorum.

Control of this disease is based on a blood testing technique. Samples of blood, collected from each member of an adult flock, are tested against an antigen of pullorum organisms for agglutinizing antibodies. Positive reactors to the blood test are judged to be disease carriers and therefore are removed from the flock and slaughtered.

The blood agglutination test, carried out in most states for the past three or four decades, has now reduced the disease to one of relatively

[1] Journal Paper No. J-4047 of the Iowa Agricultural and Home Economics Experiment Station, Ames, Iowa. Project No. 1326.

[2] Department of Poultry Husbandry. [3] Department of Statistics. [4] Department of Statistics; now at System Development Corporation.

minor importance. For example, the number of reactors reported by participants in the Iowa Control program was less than .07 of one percent in 1959 while in the Massachusetts program no positive reactors were reported in that year.

When a disease such as Pullorum is widely distributed but at a low incidence on a state-wide or national level, it becomes problematical whether to continue past practices of completely blood testing all birds in a flock, or whether blood testing only a sample, or even no birds at all, of each flock might be more economical.

The present paper investigates the extent to which 100 percent blood testing is justified, assuming random, that is binomial, infestation of adult flocks. Specifically, we attempt to determine the most economical flock proportion to be blood tested in any given year and locale, as a function of certain values and costs and of the prevailing binomial infestation rate.

Economic optimizations in the presence of binomial a priori distributions have been studied previously, as for example in [1], [2], [3] and [4]. However, this prior work is almost entirely concerned with industrial inspection problems. The present paper is intended in part to illustrate the fact that such methodology, initially intended for industrial application, is equally applicable in the biological realm.

We recognize that the assumption of binomial infestation might in certain instances be improved upon, by suitable contagion models for example. However, this assumption seems not unreasonable in the case of Pullorum; in this case, the adult flock is composed of birds which, though able to transmit the disease to their progeny through the egg, have low contagious influence as carriers.

2. ALTERNATIVE FLOCK TESTING POLICIES; VALUES AND COSTS

We consider the following one-parameter (the parameter is n) family of flock testing policies:

> Sample and blood test n birds out of a flock of N. If the sample contains no reactors (reaction assumed equivalent to infection), return the n birds to the flock and blood test no more. If the sample contains one or more reactors, blood test the entire flock; slaughter all discovered reactors.

The relative economic worth of each of these $N + 1$ competing flock testing policies are now evaluated in the light of the following values and costs.

i: blood test cost per bird,
c: carcass value per bird,
$a(k)$: average value of a member of a flock of size N that contains k undiscovered reactors,
$b(k)$: value of a non-reactor from a completely tested flock of size N containing k discovered reactors.

Note that $a(0) = b(0)$ is the value of a member of a flock containing no reactors. Again, the value of a discovered reactor from a completely tested flock is c. Finally, the precise shapes of the function $a(k)$ will depend on the likelihood that the progeny from a mating involving an undiscovered reactor will become acutely infected, thereby precipitating an epidemic in the progeny flock. The shape of the function $b(k)$ will relate largely to loss of good-will. If good-will is not an important factor, it will not be unreasonable to set $b(k) = b(0)$ for all k; indeed, this is the assumption made at the end of Section 5.

3. PRELIMINARY CONSIDERATIONS

Computationally tractable forms of $a(k)$ and $b(k)$ are as follows.

$$\begin{aligned} a(0) &= A,\, a(k) = \alpha \quad \text{for} \quad k \neq 0, \\ b(0) &= A,\, b(k) = \beta \quad \text{for} \quad k \neq 0. \end{aligned} \tag{1}$$

It is shown below that the most profitable n is either 0 or N. If this can, for the moment, be assumed, then the derivation of the most profitable sample size (i.e. the choice between 0 and N) becomes a matter of simple arithmetic, at least for the limiting cases of very small or very large flock size N.

Consider a large number M of birds, grouped into m flocks of $N = M/m$ birds. The case of very small flock size is typified by $N = 1$, while the case of very large flock size is typified by $m = 1$.

For $N = 1$, the average value of the M birds (i.e. one-bird flocks) equals

$$M(1-\pi)A + M\pi\alpha, \qquad \text{if} \quad n = 0, \tag{2a}$$

$$M(1-\pi)A + M\pi c - Mi, \qquad \text{if} \quad n = N\ (=1), \tag{2b}$$

so that 100 percent testing will be more (less) profitable than no testing according to whether

$$\pi(c - \alpha) > (<)\ i. \tag{3}$$

For $m = 1$, the average value of the M birds (i.e. one flock of M birds) equals

$$M(1 - \pi)\alpha + M\pi\alpha = M\alpha, \qquad \text{if} \quad n = 0, \tag{4a}$$

$$M(1 - \pi)\beta + M\pi c - Mi, \qquad \text{if} \quad n = N\ (=M), \tag{4b}$$

so that 100 percent testing will be more (less) profitable than no testing according to whether

$$\beta - \alpha > (<)\ \pi(\beta - c) + i. \tag{5}$$

Expressions (2*a*) and (2*b*) arise from the fact that, under both zero and complete testing, a proportion $(1 - \pi)$ of the M one-bird flocks will be disease-free, contributing an amount $M(1 - \pi)\ A$ to average value. In addition, there will be a value contribution from the $M\pi$ one-bird flocks containing a reactor; this contribution will amount to $M\pi\alpha$ in the case of no testing and to $M\pi c$ in the case of 100 percent testing. Finally, there is the testing cost Mi that is incurred under 100 percent testing.

Expressions (4a) and (4b) arise from the fact that a very large flock will contain an approximate proportion π of reactors; hence, under 100 percent testing, there will be approximately $M\pi$ birds contributing carcass value c to the flock, and approximately $M(1 - \pi)$ birds contributing value β. Similarly, if M is large enough to make $M\pi$ large, the flock of M birds will contain at least one reactor with probability essentially 1, leading to an average per-bird value of α in the absence of testing.

The criteria represented by (3) and (5) constitute an almost adequate solution for the case of the value assumptions given in (1). The further computations of Section 4 will serve only to validate the assertion that the most profitable sample size must either be zero or N, and will lead, as well, to the analogues of (3) and (5) for flocks of intermediate size.

4. THE COMPUTATION OF THE MOST PROFITABLE SAMPLE SIZE FOR THE CASE OF CONSTANT VALUE DIFFERENCE $b(k) - a(k)$

This is the case typified by form (1) of the functions $a(k)$ and $b(k)$. The computations proceed as follows. Let

$$\phi_k = \text{the binomial probability of } k \text{ reactors in a flock of size } N = \binom{N}{k} \pi^k (1 - \pi)^{N-k}, \tag{6}$$

and let

$h_{n,k}$ = the hypergeometric probability of obtaining n non-reactors when drawing a random sample of size n from a flock of N birds containing k reactors and

$$(N - k) \text{ non-reactors} = \binom{N-k}{n} \Big/ \binom{N}{n}. \qquad (7)$$

Then for any particular one of the $(N + 1)$ alternative policies presented in Section 2, say the policy corresponding to sample size n,

$$\Pr \{\text{number of reactors in the flock} = k; \text{ number of reactors in the sample} = 0\} = \phi_k h_{n,k} , \qquad (8)$$

and the "profit" ensuing if the number of reactors in the flock equals k and the number of reactors in the sample equals 0 is

$$Na(k) - ni. \qquad (9)$$

Hence the contribution to expected profit of the events involving no reactors in the sample equals the sum of the products of (8) and (9):

$$\sum_{k=0}^{N-n} [Na(k) - ni] \cdot \phi_k \cdot h_{n,k} . \qquad (10)$$

Again, for the same sample size n,

$$\Pr \{\text{numbers of reactors in the flock} = k; \text{ number of reactors in the sample} > 0\} = \phi_k \cdot (1 - h_{n,k}), \qquad (11)$$

and the "profit" ensuing if the number of reactors in the flock equals k and the number of reactors in the sample exceeds 0 is

$$(N - k) \cdot b(k) + kc - Ni. \qquad (12)$$

Hence the contribution to expected profit of the events involving one or more reactors in the sample equals the sum of the products of (11) and (12):

$$\sum_{k=0}^{N} ((N - k) \cdot b(k) + kc - Ni) \cdot \phi_k \cdot (1 - h_{n,k}). \qquad (13)$$

Total expected profit will equal the sum of (10) and (13), which, neglecting terms not involving n and using the fact that $h_{n,k} = 0$ for $k \geq N - n + 1$, can be written

$$\sum_{k=0}^{N-n} \{N \cdot [a(k) - b(k)] + k \cdot [b(k) - c] + (N - n) \cdot i\} \cdot \phi_k \cdot h_{n,k} . \qquad (14)$$

Expression (14) is further reducible to

$$(1 - \pi)^n \cdot \{N \cdot E[a(k) - b(k)] + E[k \cdot (b(k) - c)] + (N - n) \cdot i\}, \qquad (15)$$

where the expectation $E[\ ]$ is with respect to the chance variable k having a binomial distribution with parameters $(N - n)$ and π. This type of expectation arises from the fact that

$$\phi_k h_{n,k} = (1 - \pi)^n \left[\binom{N - n}{k} \pi^k (1 - \pi)^{N-n-k} \right].$$

Further reduction leads, except for the additive term $(1-\pi)^N \cdot N(\beta-\alpha)$, to the expression

$$(1 - \pi)^n (T - nS) \equiv P(n), \tag{16}$$

where

$$T = N[\alpha - \beta + \pi(\beta - c) + i], \tag{17}$$

$$S = \pi(\beta - c) + i > 0. \tag{18}$$

Differencing $P(n)$ now yields

$$\Delta(n) = P(n + 1) - P(n) = (1 - \pi)^n [(n\pi - 1)S - \pi(T - S)],$$

which shows that $\Delta(n)$ is negative for $n < (T/S) + (1/\pi) - 1$, and is positive for $n > (T/S) + (1/\pi) - 1$. This implies that no $P(n)$ for $0 < n < N$ can be larger than the larger of $P(0)$ and $P(N)$, which in turn implies that the most "profitable" sample size n either is 0 or N. Establishing this fact (which is reminiscent of conclusions reached in [1] and [2]) was one of the two objectives set for this section in the last paragraph of Section 3.

The second objective set for this section was to derive a condition analogous to (3) and (5) for determining the relative profitabilities of the two sample sizes 0 and N for intermediate flock size N. But this now is simply a matter of comparing $P(0)$ and $P(N)$, where $P(n)$ is given by (16). This yields the conclusion that 100 percent testing is more (less) profitable than no testing at all according to whether

$$(\beta - \alpha)[1 - (1 - \pi)^N] > (<)\ \pi(\beta - c) + i. \tag{19}$$

We note that (19) does indeed specialize to (3) and (5) for N equal, respectively, to 1 and to ∞.

Although details are outside the scope of this paper, it may be of interest to point out that condition (3) arises naturally in the computation of the sequential Bayes test policy for the value assumptions (1). Consider the much (though no doubt impractically) enlarged set of flock testing policies consisting of all sequential plans with terminal acts A and R:

A : Stop testing; collect c for every reactor culled out so far; return non-reactors to the flock; eventually collect α or β per bird of the unculled portion of the flock, depending on whether or not this portion contains at least one reactor, unless the entire flock is reactor-free, in which case collect A.

R : Test 100 percent; collect c for reactors, and β for non-reactors, unless the entire flock is reactor-free, in which case collect A.

A straightforward application of the methodology given in [2] then shows that 100 percent testing is the most economic of the policies in this enlarged set if $\pi(c - \alpha) > i$. However, the complementary prescription for no testing if $\pi(c - \alpha) < i$ does *not* apply in this case.

5. THE COMPUTATION OF THE MOST PROFITABLE SAMPLE SIZE FOR THE CASE OF LINEARLY INCREASING VALUE DIFFERENCE $b(k) - a(k)$.

A value assumption alternative to (1) is

$$\begin{aligned} a(k) &= A - \frac{k}{N}(A - \gamma), \qquad 0 \leq k \leq N, \\ b(k) &= A - \frac{k}{N}(A - \delta), \qquad 0 \leq k \leq N. \end{aligned} \tag{20}$$

Using (15) [which was derived without reference to any specific form of $a(k)$ and $b(k)$], expected profit now becomes, except for additive constants not involving n,

$$P(n) = (1 - \pi)^n \left(\frac{N - n}{N}\right)(T + nS), \tag{21}$$

where

$$T = (A - \delta)(\pi - \pi^2)(N - 1) + N[\pi(\gamma - c) + i], \tag{22}$$

$$S = (A - \delta)\pi^2 \geq 0. \tag{23}$$

The function $P(n)$ given by (21) is best described in terms of the following five parametric cases.

Case I: $S > 0$, $T \geq 0$. As n increases from $-\infty$, $P(n)$ rise steadily from $-\infty$, crosses the n-axis at $n = -T/S$, equals T at $n = 0$, turns downward somewhere between $n = -T/S$ and $n = N$, crosses the n-axis once more at $n = N$, turns upward somewhere beyond $n = N$, and approaches the n-axis asymptotically from below as n approaches $+\infty$.

Case II: $S > 0$, $0 > T > -NS$. As n increases from $-\infty$, $P(n)$ rises steadily from $-\infty$, equals T at $n = 0$, crosses the n-axis at $n = -T/S$, turns downward somewhere between $n = -T/S$ and $n = N$, crosses the n-axis once more at $n = N$, turns upward somewhere beyond $n = N$, and approaches the n-axis asymptotically from below as n approaches $+\infty$.

Case III: $S > 0$, $T \leq -NS$. As n increases from $-\infty$, $P(n)$ rises steadily from $-\infty$, equals T at $n = 0$, crosses the n-axis at $n = N$, turns downward somewhere between $n = N$ and $n = -T/S$, crosses the n-axis once more at $n = -T/S$, turns upward somewhere beyond $n = -T/S$, and approaches the n-axis asymptotically from below as n approaches $+\infty$.

Case IV $S = 0$; $T > 0$. As n increases from $-\infty$, $P(n)$ decreases steadily from $+\infty$, equals T at $n = 0$, crosses the n-axis at $n = N$, turns upward somewhere beyond $n = N$, and approaches the n-axis asymptotically from below as n approaches $+\infty$.

Case V: $S = 0$, $T \leq 0$. As n increases from $-\infty$, $P(n)$ increases steadily from $-\infty$, equals T at $n = 0$, crosses the n-axis at $n =$ N, turns downward somewhere beyond $n = N$, and approaches the n-axis asymptotically from above as n approaches $+\infty$.

Cases IV and V are easily summarized as follows: If $A = \delta$ (corresponding, as indicated at the end of Section 2, to the absence of the good-will factor), then the only policies in contention are 100 percent testing ($n = N$) and no testing ($n = 0$), and 100 percent testing will be more (less) profitable than no testing according to whether

$$\pi(c - \gamma) > (<)\, i. \tag{24}$$

It seems of interest to note the resemblance of (24) and (3).

Cases III and V are summarized by: If $T + NS \leq 0$, it is most economical to test 100 percent.

For Case II, the most economical sample size is the n between $-T/S$ and N at which $P(n)$ turns downward. This case is of interest since it establishes the possibility of an optimum sample size other than 0 or N. This possibility has already been noted in [3].

For Case I, the most economical sample size is either $n = 0$ or the n at which $P(n)$ turns downward, depending on the relative magnitudes at these two sample sizes. Note that, for $T = 0$, the n at which $P(n)$ turns downward must be greater than zero, so that, as in Case II, the most economical sample size will be other than 0 or N.

6. EXAMPLE AND CONCLUSIONS.

Consider the case when good-will is not a factor, i.e. the case of constant $b(k)$. In this case both the formulation of Section 4 and that of Section 5 imply that only the two policies of zero and 100 percent testing are in contention, the choice between these depending on the direction of a simple inequality.

Defining the *critical testing cost* i_c to be the testing cost for which zero and 100 percent testing are equally economical, it seems of interest to compute i_c for both formulations, using comparable value figures.

A typical value for a non-reactor in a fully tested flock is \$3, a typical average value of a member of an untested flock is \$2.40, and typical values for c, π and N are \$.50, 10^{-5} and 500.

Computing i_c in the spirit of Section 4, we therefore set $A = \beta = \$3$ and $\alpha = \$2.40$. Replacing $1 - (1 - \pi)^N$ by $N\pi$ (allowable since $N\pi$ is small), criterion (19) then becomes $\$3 \times 10^{-3} > (<) \$2.5 \times 10^{-5} + i$, which means that $i_c = \$0.003$, i.e. that no testing is most profitable unless the cost of testing falls below 3 mills per bird.

Computing i_c in the spirit of Section 5, we set $A = \delta = \$3$. In addition, we interpret \$2.40 to be the value of the linear function $a(k)$ evaluated at $k = N\pi$, the expected number of reactors in the flock. This implies a per-bird disaster value of $\gamma = -\$6 \times 10^4$ for an unsuspected 100 percent infected flock. Criterion (24) then becomes $\$0.6 > (<) i$, which means that $i_c = \$0.60$, i.e. that 100 percent testing is most profitable unless the cost of testing rises above 60¢ per bird.

In practice, the cost of testing is approximately seven cents per bird. Since this cost is well bracketed by the critical costs 0.3¢ and 60¢ derived above, we learn that the shape of the value function $a(k)$ [and of course also that of $b(k)$] must be determined rather accurately if the methodology presented here is to be applied.

REFERENCES

[1]. Barnard, G. A. [1954]. Sampling inspection and statistical decisions. *JRSS (B)*, *16*, 151–72.

[2]. Eisenberg, H. B. [1959]. Bayesian lot-by-lot sampling inspection. *M. S. Thesis* Iowa State University, Ames, Iowa.

[3]. Eisenberg, H. B. [1960]. Bayesian sampling inspection for binomial a priori distributions and quadratic loss functions. *Unpublished.*

[4]. Hald, A. [1960]. The compound hypergeometric distribution and a system of single sampling inspection plans based on prior distributions and costs. *Technometrics 2*, 275–340.

NUMERICAL ASPECTS OF THE REGRESSION OF OFFSPRING ON PARENT[1]

H. E. McKean and B. B. Bohren
Population Genetics Institute, Purdue University
Lafayette, Indiana, U. S. A.

INTRODUCTION

In an earlier paper (Bohren, McKean, and Yamada, [1961]) three currently employed techniques for estimating the regression of offspring on parent, and thereby heritability in the narrow sense, were compared and contrasted with respect to their efficiencies of estimation. The general conclusion, based on theoretical considerations and an empirical study of five generations of a closed poultry flock (Yamada, Bohren, and Crittenden, [1957]), was that under the circumstances considered, the method of regression of offspring means on parent's record (method 1) was inferior to the method of regression of individual offspring on parent (method 2) (with the parent's record repeated once for each of its offspring) and to (method 3) the Kempthorne-Tandon technique (Kempthorne and Tandon, [1953]).

The success of the Kempthorne-Tandon technique depends upon knowledge of a parameter ρ, the correlation between deviations of two offspring of the same parent from the predicted breeding value of the parent, and its expected superiority over the second method depends upon the magnitude of ρ. Usually ρ is guessed in the light of prior knowledge, and weights are assigned to the families according to the guessed value of ρ. In the first paper it was shown that, under the assumption of all genetic variance being additive, $\rho \leq .067$ or $\leq .079$, depending upon whether the mating structure is random or hierarchal.

The results obtained in the previous paper specifically depended upon the particular distribution of family sizes encountered in the five analyses, and upon the accuracy of the estimated values of $T = \rho/(1-\rho)$. The purposes of this paper are to consider the efficiency loss incurred by mis-guessing ρ in the Kempthorne-Tandon technique, and to investigate the factors involved in the relative efficiency of the other two methods.

THEORY

We consider a breeding experiment in which s sires are selected

[1] Journal Paper Number 1688, Purdue University Agricultural Experiment Station.

from the population, sire i being mated to a random sample of d_i distinct dams. The mating structure is then hierarchal, with the progeny of sire i having phenotypic values given by the model

$$Y_{ijk} = \mu_i + \beta(X_{ij} - \mu) + e'_{ijk} , \tag{1}$$

which is equation (11) of the previous paper.

It has been pointed out that the three methods under consideration are merely special cases of the general unbiased weighted regression estimation procedure. The difference between the methods involves only the weights applied to each progeny-group deviation: (1) $w_{ij} = 1$ for the progeny means on parents technique, (2) $w_{ij} = n_{ij}$ for the repeated parents technique, and (3) $w_{ij} = n_{ij}/(1 + n_{ij}\tau)$ for the Kempthorne-Tandon technique, where τ is a guessed value of T. If $\tau = T$, the third technique is the minimum variance technique, whereas, if $\tau = 0$, the third technique reduces to the second. Furthermore, when the family sizes are equal ($n_{i1} = n_{i2} = \cdots$, for all i), all three methods are identical in efficiency. Since most experimental data will involve unequal family sizes, interest will center on considering this situation.

The question, "When may I use method 1 with little loss in efficiency?", is a pertinent one. First, let us examine the optimum choice of weights for which the variance of the estimate of β will be minimized. These optimal weights w^*_{ij} (say), where $w^*_{ij} = n_{ij}/(1 + n_{ij}T)$, will be approximately equal (hence method 1 appropriate), for $T > 0$, under one or more of three distinct sets of circumstances:

1) All n_{ij} are large. This follows immediately from

$$\lim_{n_{ij}\to\infty} \left[\frac{n_{ij}}{1 + n_{ij}T}\right] = \frac{1}{T}.$$

2) S^2_n , the variance of the family sizes, is zero or very small.

3) T is large. This follows from the fact that

$$\lim_{\substack{T\to\infty \\ (\text{or } \rho\to 1)}} \frac{w^*_{ij}}{w^*_{i'j'}} = 1 ,$$

independent of n_{ij} and $n_{i'j'}$; thus for large T, $w^*_{ij} \doteq w^*_{i'j'}$. In view of this, method 1 may also be considered as a special case of method 3 where τ, the guessed value of T, is allowed to approach ∞.

It is of interest, therefore, to determine a readily accessible statistic which depends upon $\bar{n}$ (the average dam family size), S^2_n , and T, upon which a decision to use or not use method 1 as opposed to method 2 may be based.

It is easy to show that the coefficient of variation among the optimal weights w^*_{ij} is approximately

$$\psi(T) = \frac{S_n}{\bar{n}(1 + \bar{n}T)} = \frac{C.V.(n)}{1 + \bar{n}T},$$

where $C.V.(n) = S_n/\bar{n}$ = coefficient of variation of the dam family sizes. Since $\psi(T)$ is a function of all three of the statistics mentioned above, it is of interest to consider the relationship between relative efficiency (Min $V(\hat{\beta})/V(\hat{\beta})$) of the various techniques and $\psi(T)$ [even though $\psi(T)$ may seriously differ from the true coefficient of variation of the w^*_{ij}].

For any choice of τ, we may define

$$\psi(\tau) = C.V.(n)/(1 + \bar{n}\tau). \tag{2}$$

Clearly, $\psi(0) = C.V.(n)$, $\psi(\infty) = 0$, and $\psi(\tau)$ is a monotonically decreasing function of τ. Since the range $0 \leq \tau \leq \infty$ corresponds one-to-one to the range in ψ of $C.V.(n) \geq \psi \geq 0$, it is more convenient to express graphically a relationship between ψ, T, and relative efficiency rather than τ, T, and relative efficiency.

Given a fixed set of family sizes n_{11}, n_{12}, $\cdots$ the relative efficiency is a function of T, the true unknown parameter of the population, and τ, the guessed value. Because of the one-to-one relationship between τ and $\psi(\tau)$, the relative efficiency may be expressed as a function of T and ψ. The explicit form of this function may be obtained by solving equation (2) for τ and substituting the resulting expression in ψ into the equation for relative efficiency:

$$R(T, \tau) = \sum_{i=1}^{s} (1/\sigma^2_{\hat{\beta}_i}) \Big/ \sum_{i=1}^{s} (1/\sigma^{2*}_{\hat{\beta}_i}),$$

where $\hat{\beta}_i$ = estimated regression within the ith sire using τ as the guessed value of T,

$\hat{\beta}^*_i$ = estimated regression within the ith sire using $\tau = T$ (optimum), and

$$\sigma^2_{\hat{\beta}_i} = \sigma^2(1 - \rho) \sum_j n_{ij} \frac{1 + n_{ij}T}{(1 + n_{ij}\tau)^2} \frac{(X_{ij} - \tilde{X}_{i.})^2}{[\sum_j w_{ij}(X_{ij} - \tilde{X}_{i.})^2]^2},$$

$$\sigma^{2*}_{\hat{\beta}_i} = \sigma^2(1 - \rho) \Big/ \sum_j w^*_{ij}(X_{ij} - \tilde{X}_{i.})^2,$$

which correspond to equations (15) and (16) of the previous paper.

The graph of a typical relationship between ψ, T, and relative efficiency is given in Figure 1. The x- and y-coordinates of any point on the surface pictured represent respectively, the coefficient of varia-

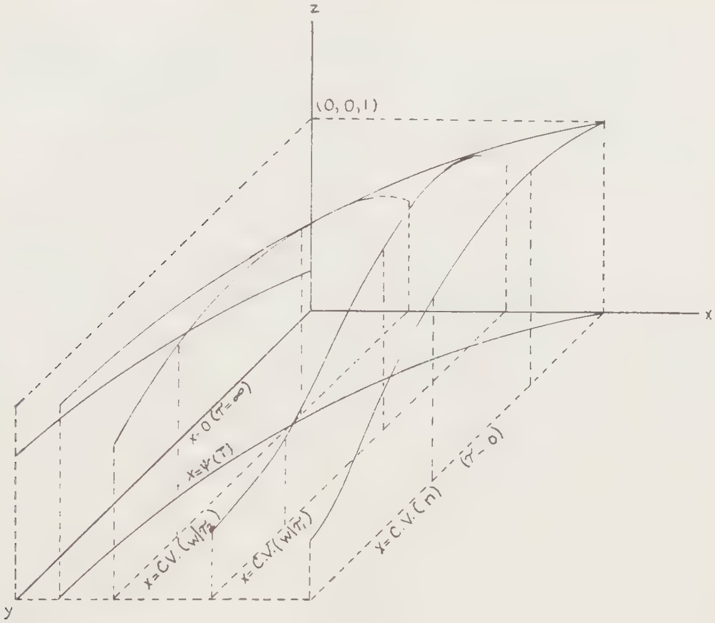


FIGURE 1.

Graphical Representation of a Typical Relationship Between $x = \psi$, $y = T$, and z = Relative Efficiency.

tion of the weights resulting from a choice of τ, and the true value of the parameter T, whereas the z-coordinate is the resultant relative efficiency. Several interesting properties should be noted:

1) The points in the (x, y)-plane corresponding to $\tau = T$ (optimal efficiency) lie on a curve whose equation is $x = \psi(T)$. The surface representing relative efficiency thus has a ridge which coincides with the projection of this curve onto the plane $z = 1$.
2) The relative efficiency falls off from this ridge in all directions.
3) The relative efficiency of method $1(\tau = \infty)$ is a strictly increasing function of T.
4) The relative efficiency of method $2(\tau = 0)$ is a strictly decreasing function of T.
5) The relative efficiency of method $3(0 < \tau < \infty)$ is strictly increasing for $0 \leq T < \tau$ and strictly decreasing for $\tau < T < \infty$.
6) For any fixed value T, relative efficiency is a strictly increasing function of τ for $0 \leq \tau < T$ and strictly decreasing for $\tau > T$. This family of curves, generated by all possible T values, expresses graphically the consequences of mis-guessing T.

7) If $C.V.(n)$ is small, the whole picture is condensed along the x-axis so that mis-guessing the value of T will not result in serious loss in efficiency. If $C.V.(n) = 0$, the surface is degenerate, and reduces to the line $(x = 0, z = 1)$. Alternatively large values of $C.V.(n)$ may result in serious efficiency loss if $| \tau - T |$ is large.

AN ILLUSTRATION

In order to consider the effect of a poor guess for ρ, we will consider the poultry population previously studied by Bohren, McKean and Yamada [1961]. Five years data, 1952–1956 inclusive, are included in the analysis. The information pertinent to the population size and structure are given in Table 1 along with the average family sizes and the coefficients of variation of n. The relative efficiencies of the three methods for the year 1952 are shown in Table 2 for a range of values of T and τ between zero and .07. This covers most of the theoretical range of ρ, under the assumption of additive genetic variance, as found by Bohren, McKean, and Yamada [1961] and includes the estimates of T obtained from these data. It is seen that in no case is the efficiency of either the repeated parent technique ($\tau = 0$) or the Kempthorne-Tandon technique less than .97. If an intermediate value of τ is chosen under the Kempthorne-Tandon technique (say .03) it is seen that the efficiency does not decline below .99. The efficiency of the means technique is seen to be consistently poorer than the other two techniques.

TABLE 1

The Family Structure of the Populations Studied.

Year	No. Sires	No. Dams	No. Pullets	Average Family Size	C.V. (Family Size)
1952	10	106	605	5.7	.74
1953	10	92	877	9.5	.43
1954	11	78	573	7.3	.41
1955	18	108	852	7.9	.47
1956	20	132	715	5.4	.54

Analogous tables were generated for the years 1953–1956. Essentially identical results were obtained for all five years analyzed. In all years the relative efficiency of the Kempthorne-Tandon technique remains above 97 percent for $T \leq .07$ and $\tau \leq .07$.

Method 1 was consistently less efficient than method 2 in all years

TABLE 2

Relative Efficiencies of the Three Methods of Estimating β Relative to the Minimum Variance Estimator of $\beta(T = \tau)$ for Various Small Values of T for 1952 Data.

T	Method 1 (Means Tech.)	Values of τ used in the KT technique (Method 3)							
		.00[2]	.01	.02	.03	.04	.05	.06	.07
.00	.5836	1.0000	.9988	.9958	.9919	.9872	.9822	.9770	.9716
.01	.6083	.9988	1.0000	.9991	.9969	.9938	.9901	.9861	.9818
.02	.6302	.9958	.9991	1.0000	.9993	.9976	.9951	.9922	.9888
.03	.6499	.9917	.9969	.9993	1.0000	.9995	.9981	.9961	.9936
.04	.6678	.9870	.9937	.9976	.9995	1.0000	.9996	.9985	.9968
.05	.6840	.9819	.9901	.9951	.9981	.9996	1.0000	.9997	.9987
.06	.6989	.9766	.9860	.9922	.9961	.9984	.9996	1.0000	.9997
.07	.7126	.9713	.9818	.9889	.9937	.9968	.9987	.9998	1.0000

[2] $\tau = 0$ is Method 2

over the range $0 \leq T \leq .07$, but was relatively more efficient than in 1952. Letting E denote the efficiency of method 1 (means), then, as T rises from zero to .07, in:

1952, E rises from .5836 to .7126,
1953, E rises from .9142 to .9615,
1954, E rises from .9046 to .9587,
1955, E rises from .9370 to .9682, and
1956, E rises from .8058 to .8703.

Table 3 gives a direct comparison of the relative efficiencies of methods 1 and 2 when large values of T are involved, which may, but do not necessarily, arise when maternal effects, non-additive genetic effects, and/or environmental correlations are present. It is interesting to note that the T-value at which the two methods are equally efficient varies considerably due to differences in the distribution of dam family sizes, varying from approximately $T = .09$ in 1953 to $T = .32$ in 1952. The corresponding values of $\psi(T)$, the coefficients of variation of the optimum weights w_{ij}^*, show much more year-to-year stability, however, ranging from a low of .23 in 1953 to a high of .27 in 1952.

Let T_0 represent the value of T for which the two methods are equally efficient. The values $\psi(T_0)$ will not ordinarily be limited to as small a range as observed here (see Table 4). It is quite possible to have much lower values of $C.V.(n)$, which is an upper bound for $\psi(T_0)$. Hence one cannot claim any invariance properties for $\psi(T_0)$. On the

TABLE 3

Relative efficiencies of methods 1 and 2 for large values of T.

Year	Method	.10	.20	.30	.40	.50
1952	1(Means)	.748	.825	.870	.900	.920
	$2(\tau = 0)$	.956	.912	.880	.856	.838
1953	1(Means)	.970	.984	.990	.993	.995
	$2(\tau = 0)$	.963	.932	.912	.898	.889
1954	1(Means)	.968	.984	.990	.993	.995
	$2(\tau = 0)$	.971	.947	.932	.921	.914
1955	1(Means)	.974	.985	.990	.993	.995
	$2(\tau = 0)$	.974	.952	.937	.926	.918
1956	1(Means)	.887	.924	.945	.958	.966
	$2(\tau = 0)$	.975	.945	.920	.901	.885

TABLE 4

Illustration of the Utility of $\psi(T)$ in Choosing Between Methods 1 and 2.

				Estimated Standard Error	
Year	$\hat{T}$	$\psi(\hat{T})$	$\psi(T_0)$	Method 1	Method 2
1952	.059	.55	.27	.121	.102
1953	.029	.34	.23	.085	.083
1954	.051	.30	.24	.152	.148
1955	.055	.33	.26	.086	.084
1956	.034	.46	.24	.108	.094

basis of information available the recommended procedure for selecting an estimation technique is as follows:

A. From previous experience, or by any other reasoning, assume a value τ for T.

B. If possible, use the Kempthorne-Tandon technique, constructing weights using this assumed value τ.

C. If it is neither possible nor convenient to use method 3, compute $\psi(\tau)$. For large values of $\psi(\tau)$ (say $> .4$) preference is for method

2 whereas for small values of $\psi(\tau)$ (say $< .1$) method 1 is preferred. For intermediate values of $\psi(\tau)$ there should be no substantial differences between methods one and two.

To illustrate the use of this approach, consider Table 4. Here, estimates of T, obtained by Bohren *et al.* [1961], for each of the years 1952–1956, are used to calculate $\psi(\hat{T})$. For comparison, $\psi(T_0)$ is also entered, together with the estimated standard errors of $\hat{\beta}$. The values $\psi(T_0)$ indicate a preference for method 2 in each case, but T_0 is obtained only after tedious calculation. The rule of thumb suggested would lead to strong preference for method 2 in 1952 and 1956 and no preference in the other years.

The results of these investigations indicate that in many circumstances a poor choice of technique can result in serious loss in efficiency in the estimation of β, and thus of heritability in the narrow sense ($h^2 = 2\beta$).

It would seem that, once the expense of experimentation and data collection has been incurred, the investigator is under an obligation to obtain highly efficient estimates. However, in some instances the loss in efficiency is negligible if one of the two standard techniques is used. The use of $\psi(\tau)$ to aid in the choice between the standard methods seems promising.

ACKNOWLEDGEMENT

The authors wish to express appreciation to Dr. Wendell Kyle, Dr. Truman Martin, and the referees for their many helpful comments on this series of two papers.

REFERENCES

1. Bohren, B. B., McKean, H. E. and Yamada, Y. [1961]. Relative efficiencies of heritability estimates based on regression of offspring on parent. *Biometrics 17*, 481–91.
2. Kempthorne, O. and Tandon, O. B. [1953]. The estimation of heritability by the regression of offspring on parent. *Biometrics 9*, 90–100.
3. Yamada, Yukio, Bohren, B. B. and Crittenden, L. B. [1957]. Genetic analysis of a White Leghorn closed flock apparently plateaued for egg production. *Poultry Science 37(3)*, 565–80.

SOME HYPOTHESES CONCERNING TWO PHASE REGRESSION LINES

P. Sprent
East Malling Research Station,
Maidstone, Kent, England.

1. INTRODUCTION

The regression of a growth measurement y on time x can often be reasonably represented by two intersecting straight lines, one being appropriate when x takes values below and the other when x takes values above a certain fixed but often unknown value corresponding to the intersection. Such regressions are here called two-phase regressions, the intersection of the phases being referred to as the *change-over point*, and the value of x at which it occurs being called the *change-over value.*

Situations in which such regressions might occur include the onset of a disease resulting in a reduced growth rate; the application of a treatment having an immediate stimulating or inhibiting effect; the occurrence of an extremely hot or cold day or some other change in external conditions; physical injury of an organism.

In a study of the compatibility of peach scions on plum rootstocks Garner and Hammond [1938] noted that the peach variety Hale's Early developed at constant but different rates on compatible and incompatible rootstock-scion unions up to a certain date. After that date the growth rate in the compatible case continued at a new constant rate, whilst in the incompatible case all growth then ceased.

Thus for a compatible union there was a typical two-phase regression, whilst for the incompatible union a rather special case occurred in which the slope of the second phase was zero. A further example is given in Section 4 in which the date of phase change in relation to time elapsed after application of treatments is of interest.

If x and y are growth measurements on two different parts of the same organism, and Huxley's allometric growth law operates (i.e. there is a linear relation between log x and log y) it has sometimes been found that sudden changes in slope occur. Reeve and Huxley [1945] have discussed this situation with reference to changes in growth equilibrium in crustacea at sexual maturity. Skellam *et al.* [1959] also refer to such phase changes in allometric growth, mentioning earlier work of Teissier on this subject.

A biologist will often postulate a two-phase linear regression rather than some alternative such as a parabolic one on largely intuitive grounds, and his decision on this point must be to some extent a matter of experience and common sense, as is generally the case in selecting appropriate hypotheses and models for statistical examination. In many cases a two-phase regression can only be a reasonable approximation, adequate for many purposes, but by no means a complete description of what is taking place.

It is likely that two-phase regressions also arise in economic and in industrial production problems.

2. HYPOTHESES

In the present context questions of the following type are likely to interest the experimenter.

A. Necessity for, and adequacy of, a two-phase fit.

B. Investigation of change-over points and values for individual regressions.

C. Relationships between lines within each phase for several regressions.

D. Relative positions of change-over points for several regressions. For example, interest may attach to one or more of the following hypotheses.

In relation to *A*:

AI. One phase adequate (i.e. a conventional linear regression suffices),

AII. A two-phase linear regression is adequate;

In relation to *B*:

BI. The observed change-over value is consistent with a specified theoretical value $x = \gamma$;

In relation to *C*:

CI. The lines within a phase are all parallel to one another,

CII. The lines within a phase all coincide;

In relation to *D*:

DI. The change-over points are collinear, all lying on a specified line $y = \alpha + \beta x$, (or $x = \gamma$ if β infinite),

DII. As for *DI* except that α (or γ) not specified,

DIII. The change-over points are collinear on an unspecified line.

In practice, anyone accepting *AI* would proceed no further in this contest. If *AI* were rejected and *AII* accepted, a test of *BI* might be required, and with more than one set of data an experimenter may wish to proceed to some of the C or D hypotheses. For instance, he might test and accept CI, reject CII as the result of a test (or without test if he considered it incompatible with his data) and then test and accept DII for the case of infinite slope. If x were a time measure, acceptance of DII with β infinite would imply that all phase changes occured at the same time, and is thus likely to be of practical interest. If one wants to know if a phase change can be associated with an event occurring at a particular x value, such as the application of a treatment, or the occurrence of a very hot day, a test of BI is required.

The above hypotheses are by no means exhaustive. Concurrence within a phase, for instance, may be of more interest than parallelism.

In this paper two assumptions that will be made are (i) the classical regression assumption that within each phase the observed y are normally distributed with mean zero and standard deviation σ about the true regression line, and (ii) that it is known between which two observed x the change-over occurs, but not precisely where between these two values it occurs.

This latter assumption calls for some comment. As will be seen, it leads to certain simple tests, and in many practical cases it is not unreasonable. Sometimes observations can only be taken at widely spaced x-values, and it may be clear that the change-over has occurred between two such observations, without a more precise statement being immediately possible.

A difficulty arises if lines are fitted to each phase making the assumption (ii), and the intersection of the fitted lines has an x co-ordinate falling outside the assumed range. The experimenter must then decide whether to attribute this to sampling errors, or an incorrect assumption, and determine the appropriate procedure accordingly. This matter is not pursued further here.

If the second assumption cannot be made because of doubt about the allocation of a small proportion of points near the change-over, the author conjectures that the methods of Quandt [1958] could be used to allocate these doubtful points without seriously upsetting the procedures suggested here. Quandt deals with the allocation problem for a two-regime y-on-x regression where the x occur in a known order with respect to a time measure t, and the phase change occurs when a certain t-value is reached. In the present case $x = t$.

This paper deals mainly with questions classified under B and D, but we first briefly discuss hypotheses involving A and C.

If one has an independent estimate of error variance, the test of AI is simple and standard, being a test for significance of deviations from regression. If deviations from regression provide the only estimate of error variance and indicate significance of linear regression, some judgement is needed to decide whether a two-phase regression (or any other fit such as a quadratic regression for example) is worth considering. The test of *AII* is likewise essentially a test of the adequacy of two linear regressions, one in each phase.

If *AII* is accepted for more than one set of data and tests for parallelism or coincidence (*CI*, *CII*) are required, these may be performed by standard methods for each phase separately.

We now consider other tests.

3. TESTS

Throughout summations over r run from 1 to 2 (corresponding to the two phases); those involving s run from 1 to n (corresponding to n sets of two-phase regressions); those involving k run from 1 to m_{rs} for the appropriate set and phase, m_{rs} being the number of observations in the rth phase of the sth set. Where it is obvious how many equations (usually 2, n or $2n$) of a given form exist, only typical ones are given.

The first test considered is that of *BI*, that is compatibility with a fixed change-over value, assuming *AII*. The test is based on the comparison of the fit by standard regression techniques of unconstrained lines in each phase, with that obtained if the lines are constrained to intersect at a point whose x co-ordinate γ is given.

Appropriate models for the first and second phases respectively are $E(y) = \alpha_1 + \beta_1 x$ and $E(y) = \alpha_2 + \beta_2 x$, where the constraint to give an intersection at $x = \gamma$ is $\alpha_2 - \alpha_1 + \gamma(\beta_2 - \beta_1) = 0$.

In view of the regression assumptions the method of least squares is appropriate, and a_1 , a_2 , b_1 , b_2 , the estimators of α_1 , α_2 , β_1 , β_2 , are to be determined so as to minimize

$$\sum_r \sum_k (y_{rk} - a_r - b_r x_{rk})^2$$

subject to

$$a_2 - a_1 + \gamma(b_2 - b_1) = 0. \tag{1}$$

Although not essential, introduction of a Lagrange multiplier λ preserves a certain symmetry, and gives rise to many expressions similar to familiar regression results, together with adjustment terms. With such a multiplier the function to be minimized becomes

$$L_1 = \sum_r \sum_k (y_{rk} - a_r - b_r x_{rk})^2 + 2\lambda[a_2 - a_1 + \gamma(b_2 - b_1)],$$

whence, computing the partial derivatives, normal equations are obtained which can easily be reduced to the form

$$\sum_k (y_{rk}x_{rk} - a_r x_{rk} - b_r x_{rk}^2) + (-1)^{r-1}\lambda\gamma = 0, \qquad (2,3)$$

$$a_r = y_{r.} - b_r x_{r.} + (-1)^{r-1}\lambda/m_r\,, \qquad (4,5)$$

where $y_{r.} = \sum_k y_{rk}/m_r$, etc., m_r being the number of observations in the rth phase. The remaining normal equation is (1). Subtracting (4) from (5) and using (1) gives

$$\lambda = w[d + b_1(x_{1.} - \gamma) - b_2(x_{2.} - \gamma)],$$

where $w = m_1m_2/(m_1 + m_2) = m_1m_2/m$, say, and $d = y_{2.} - y_{1.}$.

Substituting for a_1 , a_2 , λ as given by (4) to (6) in (2) and (3) gives, upon writing Cx_ry_r , Cx_r^2 , etc., for the ordinary corrected sums of products and squares,

$$(e.g.\ Cx_1^2 = \sum_k (x_{1k} - x_{1.})^2, \quad \text{etc.}):$$

$$[Cx_1^2 + w(x_{1.} - \gamma)^2]b_1 - w(x_{1.} - \gamma)(x_{2.} - \gamma)b_2 = Cx_1y_1 - w\,d(x_{1.} - \gamma),$$

$$- w(x_{1.} - \gamma)(x_{2.} - \gamma)b_1 + [Cx_2^2 + w(x_{2.} - \gamma)^2]b_2 = Cx_2y_2 + w\,d(x_{2.} - \gamma).$$

Having solved these equations for b_1 and b_2 it is easy to obtain values for a_1 , a_2 , λ if required from (4), (5) and (6). This constrained fit is then compared with the unconstrained fit for which of course, the slopes b_r' , $(r = 1, 2)$ are given by $b_r' = Cx_ry_r/Cx_r^2$.

If there are $m = m_1 + m_2$ observations, the sum of squares for departures from unconstrained regressions has $m - 4$ degrees of freedom, two regression coefficients and two means having been estimated. For the constrained fit, the corresponding sum of squares has $m - 3$ degrees of freedom since only three independent parameters have been estimated in view of (1). It is easily shown that the difference, with one degrees of freedom, has a sum of squares given by

$$\sum_r (b_r' - b_r)Cx_ry_r + \lambda d. \qquad (7)$$

This sum of squares may also be obtained by subtracting

$$2 \sum_r b_rCx_ry_r - \sum_r b_r^2Cx_r^2 - \lambda^2/w$$

from the unconstrained regression sum of squares. It is worth noting that the constrained lines will no longer in general pass through their respective centres of mass.

An F-test of (7) against an appropriate error variance is required. A significant value indicates a worthwhile improvement in fit if the constraint is removed, and in that event BI would normally be rejected for the γ under consideration.

The above test may be extended if there are n sets of data, and given change-over values $\gamma_1, \cdots, \gamma_n$ postulated, each one associated with a given set. The regression coefficients may be estimated separately for each set by the procedure above, using the appropriate constraint each time. If desired, a joint test of the fit compared to unconstrained fits for all sets within each phase may be made. The sum of squares (7) is computed for each set, and these are added together to give a sum of squares with n degrees of freedom which can be tested against an appropriate error variance. If $\gamma_1 = \cdots = \gamma_n$ it will be noted that this test is essentially a test of DI with β infinite.

Turning now to DII there is little difficulty in developing a test if the change-overs are assumed to lie on a line of infinite slope (i.e. to occur at the same unknown x-value) providing AII and CI are accepted (i.e. all lines within each phase are parallel to one another). It will be assumed that all lines in the first phase have the same slope b_1, and all lines in the second phase have the same slope b_2, $(b_1 \neq b_2)$. A comparison has to be made between a fit constrained so that all change-over points lie on a line $x = \gamma'$, and a fit not so constrained. (Note γ' is not specified, and must be estimated.)

The model now is

$$E(y) = \alpha_{1s} + \beta_1 x \quad \text{and} \quad E(y) = \alpha_{2s} + \beta_2 x$$

for the respective phases, where $s = 1, \cdots, n$, and there are n constraints which may be written

$$\alpha_{2s} - \alpha_{1s} + \gamma'(\beta_2 - \beta_1) = 0. \tag{8}$$

As β_1, β_2 and γ' are independent of s, it is convenient to replace (8) by the n constraints

$$\alpha_{2s} - \alpha_{1s} = \gamma,$$

the geometrical significance of which, in relation to intercepts, is obvious. The function to be minimized in order to estimate these parameters (including γ in this case) is

$$L_2 = \sum_r \sum_s \sum_k (y_{rsk} - a_{rs} - b_r x_{rsk})^2 + 2 \sum_s \lambda_s (a_{2s} - a_{1s} - c),$$

the λ_s being Lagrange multipliers. Differentiation gives rise to a set of normal equations, e.g.

$$\sum_s \sum_k (y_{rsk}x_{rsk} - a_{rs}x_{rsk} - b_r x^2_{rsk}) = 0,$$

$$a_{rs} = y_{rs.} - b_r x_{rs.} + (-1)^{r-1}\lambda_s/m_{rs} ,$$

$$\sum_s \lambda_s = 0,$$

$$a_{2s} - a_{1s} - c = 0. \tag{9}$$

These lead, after some straight-forward but tedious algebra to the equations

$$c = {}_w y_{2..} - b_{2\,w}x_{2..} - ({}_w y_{1..} - b_{1\,w}x_{1..}), \tag{10}$$

$$[\sum Cx^2_{1s} + C_w x^2_{1.}]b_1 - C_w x_{1.}x_{2.}b_2 = \sum Cx_{1s}y_{1s} - C_w x_{1.}\, d, \tag{11}$$

$$-C_w x_{1.}x_{2.}b_1 + [\sum Cx^2_{2s} + C_w x^2_{2.}]b_2 = \sum Cx_{2s}y_{2s} + C_w x_{2.}\, d, \tag{12}$$

where Cx^2_{rs} , $Cx_{rs}y_{rs}$ are the usual corrected sums of squares and products for the sth line of the rth phase, and

$$
\begin{aligned}
{}_w y_{r..} &= \sum w_s y_{rs.} / \sum w_s \, , \\
{}_w x_{r..} &= \sum w_s x_{rs.} / \sum w_s \, , \\
w_s &= m_{1s}m_{2s}/(m_{1s} + m_{2s}), \\
C_w x^2_{r.} &= \sum w_s x^2_{rs.} - {}_w x_{r..}(\sum w_s x_{rs.}), \\
C_w x_{1.}x_{2.} &= \sum w_s x_{1s.}x_{2s.} - {}_w x_{1..}(\sum w_s x_{2s.}), \\
C_w x_{1.}\, d &= \sum w_s x_{1s.}\, d_s - {}_w x_{1..}(\sum w_s\, d_s), \\
d_s &= y_{2s.} - y_{1s.} \, , \cdot \text{ etc.}
\end{aligned}
$$

In the above all summations are over s.

It is easiest to solve (11) and (12) for b_1 and b_2 , then obtain c from (10). For testing purposes it is convenient to know the λ_s , which can be obtained from the easily deducible relations

$$\lambda_s = w_s(d_s - b_2 x_{2s.} + b_1 x_{1s.} - c). \tag{13}$$

Putting $m = \sum_r \sum_s m_{rs}$, the total number of observations, then for unconstrained parallel lines the sum of squares for deviations from regression has $m - 2n - 2$ degrees of freedom. When the constraints (8) apply we fit only $n + 3$ independent parameters and thus have $m - n - 3$ degrees of freedom for deviations from regression. The difference between these two deviation sums of squares has $n - 1$ degrees of freedom, and is obtained as

$$\sum_r \sum_s (b'_r - b_r)Cx_{rs}y_{rs} + \sum_s \lambda_s\, d_s \, , \tag{14}$$

where b_1', b_2' are the regression coefficients for unconstrained parallel lines.

Equations (11) and (12) simplify somewhat if the x's are the same for all sets of data.

4. EXAMPLE

R. M. Fulford (personal communication) has made counts of the number of nodes in terminal buds of spurs of the apple variety Miller's seedling from the first or outermost budscale to the youngest primordium at intervals after the application of three defoliation treatments. The treatments were applied on different dates, and the counts were made at intervals of seven or more days throughout the 1957 growing season. On each counting date at least six buds from trees that had received a given treatment were examined. The counting dates were not always the same for all treatments, but were generally close together and covered a comparable time span.

The regression of y, the number of nodes, upon x, the number of days after May 24th, was found to be consistent with the hypothesis of three sets of two-phase parallel regressions (i.e. *AII* and *CI* were satisfied). This was tested by standard methods which are not reproduced here. It appeared from the data that the change-overs might well have occurred on the same calendar date, rather than at, say, the same fixed number of days after the application of each treatment. In each phase of each regression there were observations for between six and eleven distinct counting dates x.

The raw data are not reproduced here; but relevant means, sums of squares and products, weights, etc., are given.

The content of Table 1 was computed directly from the data, and that of Table 2 from that of Table 1.

The coefficients in equations (10) to (12) are easily got from these tables, and give

$$c = 8\cdot260 - 97\cdot856b_2 + 39\cdot578b_1 ,$$

$$56926\cdot21b_1 - 777\cdot54b_2 = 11388\cdot08,$$

$$-777\cdot54b_1 + 19057\cdot96b_2 = 954\cdot93,$$

whence $b_1 = 0.2008$, $b_2 = 0.0583$, $c = 10.502$. From (13) we then get

$$\lambda_1 = -4\cdot404, \qquad \lambda_2 = -5\cdot611, \qquad \lambda_3 = 10\cdot017.$$

The λ_s, apart from rounding errors, sum to zero, providing a useful check. Computing (14) we find it comes to 7.32. This sum of squares has 2 degrees of freedom. An appropriate error mean square is provided

TABLE 1.

MEANS, SQUARES, PRODUCTS, WEIGHTS

s	$x_{1s.}$	$x_{2s.}$	$y_{1s.}$	$y_{2s.}$	d_s	Cx_{1s}^2	Cx_{2s}^2	$Cx_{1s}y_{1s}$	$Cx_{2s}y_{2s}$	w_s
1	35.27	94.50	16.62	25.36	8.74	31345.09	5145.00	6335.82	360.50	23.29
2	38.62	98.00	16.48	24.69	8.21	12059.25	8232.00	2488.62	455.00	22.40
3	45.50	101.50	17.04	24.81	7.77	12348.00	5145.00	2457.00	213.50	20.57
					Sums	55752.34	18522.00	11281.44	1029.00	66.26

TABLE 2.

COMPUTATIONS REQUIRED FOR COEFFICIENTS IN NORMAL EQUATIONS

s	$w_s x_{1s.}$	$w_s x_{2s.}$	$w_s d_s$	$w_s x_{1s.}^2$	$w_s x_{2s.}^2$	$w_s x_{1s.} x_{2s.}$	$w_s x_{1s.} d_s$	$w_s x_{2s.} d_s$
1	821.44	2200.90	203.55	28972.19	207985.05	77626.08	7179.21	19235.87
2	865.09	2195.20	183.90	33409.78	215129.60	84778.82	7102.39	18022.20
3	935.94	2087.86	159.83	42585.27	211917.79	94997.91	7272.27	16222.67
sums	2622.47	6483.96	547.28	104967.24	635032.44	257402.81	21553.87	53480.74
wt. mean	39.578	97.856	8.260					
Weighted correction				103793.37	634496.48	256625.27	21660.51	53554.81
Weighted corrected *SS* or *SP*				1173.87	535.96	777.54	−106.64	−74.07

by that between counts on the same date within treatments and has the value 1.08 with 230 degrees of freedom. This routine computation is not shown. The appropriate F is thus $3.66/1.08 = 3.39$, and this is significant at the 0.05 probability level. Thus there is evidence for rejecting the hypothesis of a common change-over date, although it was clear from the data that the change-overs occurred nearer to a common calendar date than they did to a constant time lapse after application of each treatment. In the computations, particularly in calculating (14), care is needed to avoid excessive rounding errors.

5. RELATION TO FIELLER'S THEOREM

It is evident that a test of consistency of the observed change-over value with a specified $x = \gamma$, (hypothesis BI) is equivalent to a test about the intersection of two regression lines, and the test developed in Section 3 might be expected to be related to Fieller's [1940, 1944] theorem on fiducial limits of a ratio. Fieller has shown that the fiducial limits for $\theta = y/x$, where the errors in y and x are normally distributed with zero mean, are given by the roots of the quadratic in θ

$$(y^2 - t^2 v_{yy}) - 2\theta(xy - t^2 v_{xy}) + \theta^2(x^2 - t^2 v_{xx}) = 0,$$

where v_{yy}, v_{xx}, v_{xy} are estimates of the variances and covariance of y and x, each based on n degrees of freedom, and t is the appropriate value of Student's t for the required fiducial probability with n degrees of freedom.

Applying the theorem in the situation envisaged in BI, Fieller's $y = a_2' - a_1'$, and his $x = b_1' - b_2'$, in terms of the parameters of unconstrained regression lines. The estimated variances and covariances of these, using an appropriate error mean square s^2 say, are easily obtained. On the criterion of Section 3 for testing BI, a chosen γ will just be significant if

$$\sum_r (b_r' - b_r)Cx_r y_r + \lambda d - t^2 s^2 = 0. \qquad (15)$$

From Fieller's theorem it follows that such a γ is given as a root of the quadratic in γ,

$$[\gamma(b_1' - b_2') - (a_2' - a_1')]^2 - t^2 s^2 \cdot \left[\frac{1}{w} + \frac{(x_{1} - \gamma)^2}{Cx_1^2} + \frac{(x_{2.} - \gamma)^2}{Cx_2^2}\right] = 0. \qquad (16)$$

The left-hand side of (15) can be shown to be a quadratic in γ, and thus (15) and (16) will be equivalent if it can be shown that their left-hand sides are identical for three distinct values of γ. This is easily done when $\gamma = x_{1.}$ or $x_{2.}$ or $(a_2' - a_1')/(b_1' - b_2')$.

Thus using Fieller's theorem to establish fiducial limits for γ provides a set of γ's (i.e. those between the limits) which would be acceptable under the hypothesis test put forward here.

With either Fieller's theorem, or the approach used here, difficulties arise if γ is infinite, but in practice this causes no trouble in the present context as an infinite γ would imply the identity of the two phases, i.e. acceptance of *AI*, or alternatively a pair of parallel regressions, rather than the abrupt change of slope implied by γ finite.

Whilst Fieller's theorem does not link directly with the test already considered for a common unspecified change-over date for all of n sets of data, it does provide a solution to a closely related problem, namely:

Assuming there is a common-change over date, what are the p percent fiducial limits for this? Equations (10) to (12) provide estimates c, b_1, b_2, of γ, β_1, β_2 respectively. The γ' of (8) is related to γ by the expression

$$\gamma = (\beta_1 - \beta_2)\gamma',$$

thus γ' may be estimated as

$$c' = c/(b_1 - b_2). \tag{17}$$

If the value of c given by equation (10) is put in (17) and the variances and covariance of numerator and denominator obtained in the usual way, Fieller's theorem may be applied.

6. DISCUSSION

An important characteristic of the tests derived in Section 3 is that they involve linear constraints upon the parameters. This would not be the case for instance if one wished to test *DII* with β infinite but without the restriction to parallelism introduced in Section 3. Then the n constraints would become

$$a_{2s} - a_{1s} = c(b_{1s} - b_{2s}),$$

all terms being unknown. Further, if one wished to test *DII* with β finite, and no restriction to parallelism, the constraints would take the form

$$b_{1s}a_{2s} - b_{2s}a_{1s} = \alpha(b_{1s} - b_{2s}) + \beta(a_{2s} - a_{1s}).$$

These do not easily reduce to a workable system of linear constraints and we are consequently led to non-linear normal equations. It is seldom that such equations can be easily reduced to a linear set, and iterative solutions or other special methods are required. Similar difficulties arise with many hypotheses associated with problems of type *D*.

An alternative approach to questions of linearity of the change-over points might be via the fitting of lines to the observed unconstrained change-over points. As these will almost certainly be determined with different accuracies for different sets of data depending upon the numbers of points, and their positionings, in each phase, some weighting process is indicated, and the approach is entirely different from that adopted here.

Finally, from the biological viewpoint it must be emphasised that the ability to use techniques of the type suggested here in no way reduces the need for the experimenter to consider the biological reasonableness of the model he uses, and whether it is more than a convenient approximation.

7. ACKNOWLEDGEMENTS

I am indebted to R. M. Fulford for making his data available for the worked example, to F. W. M. Llewelyn for a useful suggestion about the handling of the data which inspired this investigation, and to S. C. Pearce for helpful comments.

REFERENCES

Fieller, E. C. [1940]. The biological standardization of insulin. *J. Roy. Stat. Soc. Suppt. 7*, 1–64.

Fieller, E. C. [1944]. A fundamental formula in the statistics of biological assay, and some applications. *Quart. J. Pharm. 17*, 117–23.

Garner, R. J. and Hammond, D. H. [1938]. Studies in incompatibility of stock and scion. II. The relation between time of budding and stock scion incompatibility. *Report of E. Malling Res. Sta. for 1937*, 154–7.

Quandt, R. E. [1958]. The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Stat. Assoc. 53*, 873–80.

Reeves, E. C. R. and Huxley, J. S. [1945]. Some problems in the study of allometric growth. *Essays on Growth and Form.* (Ed. Clark and Medewar). Oxford University Press.

Skellam, J. G., Brian, M. V. and Proctor, J. R. [1959]. The simultaneous growth of interacting systems. *Acta Biotheoretica 13*, (2/3), 131–44.

QUERIES AND NOTES

D. J. Finney, *Editor*

165 NOTE: A Practical Application of a Theoretically Inefficient Design[1]

C. P. Cox[2]

National Institute for Research in Dairying
University of Reading
Reading, England

1. INTRODUCTION

In experiments on milking machine techniques for reducing the time taken to milk each cow, it was desired to compare milking times given by combinations of two pulsation ratios, P_1 and P_2 with two levels, V_1 and V_2, of vacuum in the machine line. It was expected (see Section 3) that the animal to animal variability would differ substantially for the different treatments, in view of which, and the short time available for the experiment, it was decided to limit the experiment by having only two treatments per animal. The main restriction determining the design arose because the treatment combination P_1V_1 was of such special interest that it was desired to obtain an estimate of the corresponding milking time from all the available animals. The simple solution used nicely illustrates the dual features of estimation and discrimination which often interplay in biological operational research.

2. DESIGN PRINCIPLES

Each animal formed an incomplete block receiving two of the four possible treatment combinations with four successive observations on each. The special feature is that, to provide the estimate required, all animals received the combination P_1V_1 and practical considerations made it necessary for this to be the treatment for each animal in the first period. In theory therefore, the treatment differences would be confounded with any period differences but, in this case, previous experience had shown that these could safely be discounted in such a

[1] N. I. R. D. Paper No. 2462.

[2] Present address: Statistical Laboratory, Iowa State University, Ames, Iowa.

short experiment. It was logical therefore to allot the three remaining treatment combinations equally, and randomly, to the animals in the second period. Thus, if P_1V_1 is regarded simply as a control, the design itself is that properly deprecated by Yates [1936] for cases when treatment differences only are required, because of its inefficiency relative to the arrangement using all the six possible pairs. This latter design was inadmissible here because of the primary estimation requirement and it will appear that, because every observed difference contributes to each, the main effects of treatments and the interaction are estimated with sufficient accuracy to give discrimination at a useful practical level.

3. ANALYSIS

Let d_{ij} be the difference between the two period means for the jth animal on the treatment arrangement distinguished by the suffix i, as follows:

period	1		2
$i = 1$	P_1V_1	$-$	P_2V_1
$i = 2$	P_1V_1	$-$	P_1V_2
$i = 3$	P_1V_1	$-$	P_2V_2

and $j = 1, \cdots, n$ for each i.

Let $d_{i.}$ be the mean of d_{ij} over the n animals receiving the ith arrangement. Then the treatment comparisons are estimated as:

	$d_{1.}/2$	$d_{2.}/2$	$d_{3.}/2$
main effect of $P = \frac{1}{2}(P_2 - P_1)(V_1 + V_2)$	-1	1	-1
main effect of $V = \frac{1}{2}(P_1 + P_2)(V_2 - V_1)$	1	-1	-1
interaction $PV = \frac{1}{2}(P_2 - P_1)(V_2 - V_1)$	1	1	-1

Hence, if σ_i^2 is the variance of an individual d_{ij}, each effect and the interaction is estimated with variance $(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)/4n$.

With n animals in each of the three i-groups separate analyses of the d_{ij} are obtained for each group according to the simple partition:

treatment difference	1
between animals	$(n - 1)$

to which a within animals term may be added if required.

In the example, the mean differences $d_{i.}$ together with the between-animal mean squares with $n = 8$ were:

treatment difference	$d_{i.}$	mean square (7 d.f.)
$P_1V_1 - P_2V_1$	−0.244	0.1467
$P_1V_1 - P_1V_2$	−0.638	1.5225
$P_1V_1 - P_2V_2$	−1.631	1.9339

We thus obtain the main effects and interaction, in minutes, as

$$P = \tfrac{1}{2}(0.244 - 0.638 + 1.631) = 0.619,$$

and similarly,

$$V = 1.012, \qquad PV = 0.375.$$

The appreciable differences between the above three mean squares are noteworthy and were expected because milking times are functions both of the machine variables and their interactions with highly individual characteristics such as the amounts of milk secreted, neurohormonal responses and anatomical details of teats and udders.

The variance of each main effect and of interaction PV is calculated as

$$(0.1467 + 1.5225 + 1.9339)/32$$

giving a standard error of 0.336 minutes.

Since each mean square has the same number of degrees of freedom, the ordinary t-value for seven d.f. can be used for an approximate test of significance (after Cochran and Cox, [1957]) showing here that only the V-effect is significant at the five percent level.

4. DISCRIMINATION VERSUS ESTIMATION

It has been pointed out that the estimation requirement is fulfilled at the expense of the accuracy achieved for the treatment comparisons. Had a design employing all pairs of treatment combinations been applicable there would have been three more treatment groups,

$$i = 4 \quad P_2V_1 - P_1V_2 \ ,$$

$$i = 5 \quad P_2V_1 - P_2V_2 \ ,$$

$$i = 6 \quad P_1V_2 - P_2V \ .$$

The main effect of P, for example, calculated without recovery of inter-animal information, is then,

$$P = -\frac{1}{4}(d_{1.} + d_{3.} - d_{4.} + d_{6.}),$$

and, instead of n, there will now be $n/2$ animals in each group.

The variance heterogeneity in the particular example complicates the direct comparison but, if in a general case, transforming if necessary, we assume that each d_{ij} has variance σ^2, the variance of a treatment comparison is now

$$\frac{4}{16}\cdot\frac{2\sigma^2}{n} = \sigma^2/2n,$$

which compares with $3\sigma^2/4n$ obtained when only three differences are directly investigated.

Both designs provide unbiassed estimates of the milking time on P_1V_1 and the above loss of discrimination is compensated by the fact that this mean milking time is estimated with only half the variance which would have obtained if all the paired differences had been used.

Finally, with reference to the particular example, when not only the mean but the distribution of milking times on the P_1V_1 combination is of interest, the larger numbers of observations obtained with the design used will be additionally advantageous.

5. ACKNOWLEDGEMENTS

I am grateful to F. H. Dodd and M. C. Cowhig for providing the data of this example.

6. REFERENCES

Cochran, W. G. and Cox, G. M. [1957]. *Experimental Designs.* John Wiley & Sons, New York.

Yates, F. [1936]. Incomplete randomized blocks. *Ann. Eugen. 7*, 121–40.

166 QUERY: On a Graphical Sequential Test

Bross [1952] describes a graphical sequential test, and among the plotting rules is the instruction that "if new and old treatments lead to the same outcome then no information about superiority is obtained and nothing is plotted on the chart". From his chart therefore distinguishable results are considered significant whether they are 6 out of 6, or 6 out of 100 in the other 94 of which both treatments gave the same result. If one treatment kills 100 patients, and the other kills

94 and cures 6, normal probability ideas seem to suggest that there is only a 6 percent chance that there is some difference between the treatments, and not that it is less than 5 percent that there is no difference. Is it wise to ignore some of the experimental results? Should the instruction be reworded "if the fact that the outcome of the new and old treatments is the same provides no information towards the particular problem being investigated, nothing is plotted on the chart: if it supports a hypothesis that the treatments are not different, then the experiment is plotted in the square adjoining the last plot diagonally away from the origin"?

REFERENCE

Bross, I. D. J., [1952]. Sequential medical plans, *Biometrics 8*, 188.

ANSWER:

Although the query concerns a sequential procedure it applies equally well to the fixed sample procedure—the Sign Test—which is the parent of this sequential procedure so let us look at the simpler example. Both the Sign Test and its sequential analogue would be based on a study plan where one member of each patient *pair* is randomly allocated to treatment A or B (the second member receiving the remaining treatment). When both patients in the pair have the same response (both live, both die), the "tie" is ignored insofar as the significance test is concerned. Thus in the example of the query the test would be based on the six pairs where the patient on treatment A survived and the patient on B dies (and the zero cases with the opposite results). The Mosteller-Corrected Sign Test would be $(6-1)^2/6=4.17$. This exceeds 3.84 so the results would be significant at the 5 percent level.

Is it wise to ignore some of the experimental results (eg. the "ties")? Let me make it clear that the "ties" would *not* be ignored in other phases of the analysis which deal with the *clinical importance* of the difference between treatments. The question, then, is whether "ties" should be ignored in dealing with the specific point: Is there *any* difference in the responses to the two treatments? There are two technical approaches which will provide an answer to the question. First we might ask: Does ignoring the ties impair the efficiency of the procedure? The technical answer is that the loss of efficiency is negligible. Second we might ask: If significance tests are set up which involve the tied observations, will they be better than the Sign Test? The technical answer is that there is no advantage to the alternative procedures and sometimes they result in a loss of power. While both technical approaches lead to the answer "ties can be ignored" they may not satisfy an investigator who would like a simple rationale rather than a formal proof.

Therefore I will indicate a rationale for the clinical situation. The purpose of therapeutic intervention is to *change* the natural clinical course of a disease in a favorable direction. Let us suppose treatment B fails to change the natural course (i.e. to death) while treatment A occasionally can change this course. Then the *number of changes* achieved on treatment A constitute the evidence of a difference between the two treatments. If there are 6 changes, we have the same amount of evidence of a difference whether it comes from 6 pairs or 600 pairs. If there were some spontaneous remissions (eg. not due to treatment), then we would have to compare the number of changes in the two series but *this* information' would constitute the available evidence of a difference between the two treatments (irrespective of the number of pairs). When our interest lies in change of status then the "ties" are irrelevant.

I. D. J. Bross
Roswell Park Memorial Institute
Buffalo, New York, U. S. A.

167 NOTE: A Simple Method of Fitting the Regression Curve $y = \alpha + \delta x + \beta\rho^x$

B. K. Shah
Department of Statistics
University of Baroda, Baroda, India.

Patterson [1956] described a simple method for fitting an asymptotic curve

$$y = \alpha + \beta\rho^x, \qquad x = 0, 1, \cdots, n-1, \tag{1}$$

for $n = 4$, 5, 6 and 7 equally spaced ordinates. In this method ρ is estimated by r, the ratio of two linear functions of y's and α and β can be obtained by the linear regression of y on r^x.

This method can be extended to the problem of fitting the curve

$$y = \alpha + \delta x + \beta \rho^x \tag{2}$$

for equally spaced ordinates. Here also the estimate of r is a ratio of two linear contrasts of the y's and the estimates α, δ and β are calculated from a multiple linear regression of y on x and r^x.

As four parameters are to be estimated, the above method gives an exact fit for $n = 4$ using the formula

$$r = \frac{y_3 - 2y_2 + y_1}{y_2 - 2y_1 + y_0}. \tag{3}$$

The following formulae are suggested for $n = 5$, 6, 7 and 8:

$$r = \frac{5y_4 - 4y_3 - 7y_2 + 6y_1}{5y_3 - 4y_2 - 7y_1 + 6y_0}, \quad \text{for} \quad n = 5, \tag{4}$$

$$r = \frac{10y_5 - y_4 - 13y_3 - 11y_2 + 15y_1}{10y_4 - y_3 - 13y_2 - 11y_1 + 15y_0}, \quad \text{for} \quad n = 6, \tag{5}$$

$$r = \frac{10y_6 + y_5 - 6y_4 - 14y_3 - 8y_2 + 17y_1}{10y_5 + y_4 - 6y_3 - 14y_2 - 8y_1 + 17y_0}, \quad \text{for} \quad n = 7, \tag{6}$$

and

$$r = \frac{10y_7 + 5y_6 - 5y_5 - 12y_4 - 11y_3 - 7y_2 + 20y_1}{10y_6 + 5y_5 - 5y_4 - 12y_3 - 11y_2 - 7y_1 + 20y_0}, \text{ for } n = 8. \tag{7}$$

Then four formulae are chosen to give a reasonably high efficiency for independent y's with equal variances. The percentage efficiencies for $n = 5$, 6, 7 and 8, calculated from large sample formula for the variances of r and the efficient estimate (Shah and Patel, [1960]), are:

ρ	$n = 5$	$n = 6$	$n = 7$	$n = 8$
0.0	95.2	91.3	88.4	86.4
0.1	98.4	97.3	96.0	94.4
0.2	99.7	99.7	99.0	97.8
0.3	99.9	99.7	98.4	96.9
0.4	99.6	98.4	95.5	93.0
0.5	99.0	96.1	91.4	87.5
0.6	98.1	93.7	87.1	82.6
0.7	97.3	91.4	83.0	76.1
0.8	96.3	89.4	79.1	71.8
0.9	95.4	86.7	75.9	68.2

Thus the efficiencies of the proposed estimates are of about the same order as the efficiencies of Pattersons estimates of ρ for equation (1).

REFERENCES

Patterson, H. D. [1956]. A simple method for fitting an asymptotic regression curve. *Biometrics 12*. 323–29.

Shan, B. K. and Patel, I. R. [1960]. The least square estimates of the constants for the Makeham second modification of Gompertz's Law. *Jour. Maharaja Sayojirao University of Baroda 1*, Vol. IX. No. 2. 1–10.

168 QUERY: On a Follow-up Study of the Growth of Children

In conducting a long term follow-up study of the growth and development of a sample of children, the question arises as to whether losses between birth and 4 years were such as to bias the remaining sample. Two specific questions may be asked:

(a) Are losses independent of social class, age of parent etc? Leavers and stayers were compared, on several variables, using the Brandt-Snedecor method for a 2 × k table.

Example

Mother's age at child's birth	Initial Sample	Leavers	Stayers (X)	Proportion remaining (p)	pX
up to 24	49	18	31	.6327	19.6137
25–32	117	25	92	.7863	72.3396
33+	55	6	49	.8909	43.6541
Totals	221	49	172	$\bar{p}$ = .7783	$\sum pX$ = 135.6074
					$\bar{p} \sum X$ = 133.8676
					diff. = 1.7398

$$\chi^2 = \left(\sum pX - \bar{p} \sum X\right)/\bar{p}\bar{q} = 1.7398/.1725$$

$$= 10.09 \qquad (.01 > P > .001).$$

With several parental and social variables, χ^2 indicated significant differences between leavers and stayers.

(b) What is the probability that the remaining sample (stayers) could arise by random sampling from the original sample?

This would involve not a test of independence between staying and mothers' age, but a test of the difference between observed frequencies among stayers, and expected frequencies based on the proportions in the original sample.

Can χ^2 be used to answer this question? It would seem illegitimate, in so far as expected values would themselves be derived from sample values, and not from population values.

Douglas and Blomfield (1) employed a χ^2-test in this way, but in their case the initial sample consisted of all children born in a particular week in the majority of local authority areas of England and Wales. They derived expected frequencies in social class categories from the proportions in this sample, and applied them to the stayers. Is this legitimate, in so far as their expected frequencies might be said to be derived from population values?

If the χ^2-technique is inapplicable in my case, is there any other way of assessing how unrepresentative the remaining sample has become?

REFERENCE

Douglas, J. W. B. and Blomfield, J. M. [1956]. The reliability of longitudinal surveys. *Milbank Mem. Fund Quart. 34*, 227–252.

ANSWER:

Since the original sample was random, the two questions, "Are the proportions of 'leavers' and 'stayers' effectively the same in the different rows?", and "Can the sample of stayers be regarded as a random sample of the original population?" are equivalent. The χ^2-test answers the question "Could the samples in all rows (or all columns) of the table have arisen by random sampling from a single population?", which is a slightly more specific version of the first question above. Since for your data χ^2 is significant, we are led to conclude that there is a genuine difference in the proportions of mothers of different ages in the sub-populations of 'all stayers' and 'all leavers' (i.e. of all mothers in the population who would have been classified under these headings had they been included in the sample): since the original sample was random, the sample of stayers must be regarded as having been drawn at random from the sub-population of all stayers, and so cannot be regarded as randomly drawn from the whole population. No independent test of this is necessary, or indeed possible: since the two questions are logically equivalent, the same test must be used to answer both. Admittedly it appears at first sight that a different test could be derived by comparing the observed numbers of stayers with their expectations, using the statistic

$$\chi^2 = \sum \frac{[X_i - E(X_i)]^2}{\text{Var}\,(X_i)}\,, \tag{1}$$

where X_i is the number of stayers in the ith class, and the expectation and variance of X_i are calculated on an appropriate null hypothesis. This statistic does not appear to depend on the numbers of leavers: however, since the true population proportions are here unknown, this test can only be based on the null hypothesis that the X_i have arisen by independent random binomial sampling from the small numbers n_i of the original sample, with a common probability estimated by $\bar{p}$. Hence Var (X_i) must be estimated by $n_i\bar{p}\bar{q}$, so that formula (1) yields

$$\chi^2 = \sum \frac{[X_i - n_i\bar{p}]^2}{n_i\bar{p}\bar{q}}\,,$$

which, after a little manipulation, reduces to precisely the form that you have already used to calculate χ^2. Only if all the values of p are very small, so that $\bar{q}$ can be taken as effectively equal to 1, can formula (1) be approximated by the expression

$$\chi^2 = \sum \frac{[X_i - E(X_i)]^2}{E(X_i)}. \tag{2}$$

For your data $\bar{q} = 0.2217$, so that the use of formula (2) would be illegitimate, as you conjecture, and would give a χ^2 less than one-quarter the size of the correctly calculated value. As far as I am able to judge from the information given in their paper, the test used by Douglas and Blomfield is in fact invalid for this reason.

Having once decided that leavers and stayers do differ we *may* ask, quite legitimately, the further question, "Is the non-randomness of the stayers, regarded as a sample of the original population, sufficiently serious to produce large biases in the estimates to be calculated from the sample?" If the overall proportion of leavers were small, the disturbances introduced by any variation in rate of loss might well be negligible in comparison with the original sampling variation, so that the sample of stayers could be treated as though it were a random sample from the population: the resultant estimates would be biased, but the bias would be unimportant. If the true proportions in the various classes in the population are known, χ^2 may be calculated from formula (2): it must not be regarded as a test statistic, since the appropriate null hypothesis has already been rejected, but its size may be used as a criterion of the degree of non-randomness of the sample of stayers.

If the population proportions are unknown, as in your example, some indication of the seriousness of the risks involved in treating the sample of stayers as random may be obtained by considering the changes produced by leaving in the proportions of the sample falling in the various groups. For example, in the results quoted by Douglas and Blomfield the largest change in any employment group was in that of 'manual workers', which comprised 0.691 of the original sample, and 0.696 of the sample of stayers. Clearly, even though the authors were in error in maintaining that the sample of stayers could be regarded strictly as a random sample from the population, the degree of bias introduced by so regarding it was trivial. In your data, on the other hand, the proportion of mothers under 25 has fallen from 0.222 in the original sample to 0.180 in the sample of stayers, a much more serious change.

Biases resulting merely from the changes in proportions can be largely avoided, in theory, by calculating separate estimates for the different groups, and combining them according to the proportions of the original sample. However, a breakdown of the sample taking into account all criteria that show significant changes in proportions may well leave you with groups too small for satisfactory estimation, and this method cannot allow for the further danger that the leavers will almost certainly be non-typical of the population even within groups, so that a considerable bias may remain even though the effects of disproportionality are successfully removed.

M. R. Sampford
A. R. C. Unit of Statistics,
Aberdeen, Scotland.

BOOK REVIEWS

J. G. SKELLAM, *Editor*

Members are Invited to Suggest Books for Listing or Review to the Editor

16 DE JONGE, H. (editor). **Quantitative Methods in Pharmacology.** Amsterdam: North-Holland Publishing Company, 1961. Pp. xx + 391. £4.4s.

D. J. FINNEY, *University of Aberdeen, Scotland.*

In May 1960, the Netherlands Region of the Biometric Society organized a Symposium for the Society. This volume contains the texts of the 26 papers presented, together with summaries of discussions and of remarks by chairmen. The papers differ greatly in interest and importance, and any selection for comment here may reflect the taste of the reviewer as much as the quality and balance of the Symposium.

The first section of the book contains papers on sequential trials of drugs. HAJNAL presents an excellent account of sequential tests of four analgesics against a control, where assessment is in terms of a continuous variate and estimates of effects are wanted. JOHNSON introduces an important subject in a short paper on how to choose a sequential procedure that will meet practical requirements. RÜMKE suggests uses of sequential techniques for estimating doses to be used in a definitive experiment; his idea is incompletely worked out, and questions of efficiency are not discussed, but it may give rise to improvements in the strategy of using pilot experiments.

The second section, on drug standardization, includes a theoretical paper by CLARINGBOLD and EMMENS, consisting of a formal account of computational procedures appropriate to multivariate quantal responses under a general transformation, of which the normal equivalent deviate is a familiar example. Numerical illustration might have been helpful. Additional evidence is given that many types of data are insensitive to the choice of transformation, and the angle transform is therefore commended for general use. VAN STRIK's suggestion of using rank as a metameter for semi-quantitative responses may sometimes be helpful, but the method fails completely in two special cases and one may suspect that other rank orders close to these extremes will give trouble.

Evidently the most controversial session of the Symposium was that on non-parametric methods. HEMELRIJK and VAN DER VAART contribute papers on the efficiency and robustness of Wilcoxon's test; both are of considerable theoretical interest, though neither recognizes that the interpretation of numerical observations can seldom be completed solely by appeal to tests of significance. VAN EEDEN's ingenious but misleadingly titled paper illustrates this well; her method of estimating some points of the distribution function for a tolerance distribution, without assumption of any specified functional form, can evidently be extended to provide a method of estimating relative potency from two series of records which will work for some data, though for others it may be ambiguous. At best, it must be indeterminate to the extent of unit interval on the dose scale, and the author's hopes of finding a method of assigning a measure of precision to such an estimate seems at present to lack foundation. JUVANCZ's entertainingly vigorous attack on non-parametric methods perhaps relies too much on what users of statistics have done rather than on what they ought to do, but comes closer to the needs of scientific research. In "A Visual Approach to Bio-Assay," GOLDBERG presents an ingenious system of graphs and scales that enable a close approximation to the maximum likelihood estimation of parameters for a quantal response regression to be achieved by a series of additions. The account of this, even for the single parameter exponential model, is so complicated as to make this reviewer prefer the simplicity and speed of standard maximum likelihood calculations.

The sections on drug screening and on mixtures of drugs are of consistently high standard. DUNNETT and SCHNEIDERMAN each discuss problems of evaluating alternative screening rules with emphasis on economic aspects. Dunnett considers carefully the components of cost and loss that must enter into the specification of an optimal screening programme for a large number of drugs by appropriate decision rules; Schneiderman concerns himself more with the operating characteristics of his sequential procedures for screening anticancer drugs, and includes interesting comments on the use of a well-instructed computer. BROCK and SCHNEI DER seek to rehabilitate the therapeutic index by more careful definition than is customary; consideration of optimal experimental planning for the estimation of their index might throw more light on the seriousness of the loss in precision inherent in working at 5% and 95% response rates. Alternative, but not necessarily conflicting, views on quantal responses to mixtures of drugs are discussed by ARIËNS and SIMONIS and by HEWLETT and PLACKETT. The comprehensive mathematical models of the second will have the greater appeal for the statistician; for the pharmacologist, this summary of metrical aspects of Ariën's receptor theory should prove valuable, and perhaps the conjunction of the two will stimulate an effective synthesis. DIKSTEIN's short account of a hypothesis relating chain length of primary and monoquaternary amines to excitation of guinea-pig muscle must surely be the forerunner of much work on the relation of physicochemical properties and pharmacological action. DE JONGH gives a useful and simple summary of the isobole method of representing different types of action of drugs in mixture.

Inevitably a symposium of this kind includes a group of papers that defy any classification other than "miscellaneous", but those in the present volume are not the less interesting for that. GURLAND's suggestions on bioassays for estimation of minute insecticidal residues include useful practical ideas for taking account of the masking of potency by inactive materials present in plant extracts and for the "fortification" of test extracts of very low potency so as to permit adequate assays when available quantities are severely limited. Finally, a contribution from BILLEWICZ describes a statistical study of a technique of measurement (the falling drop technique for estimating the concentration of deuterium oxide): although in itself purely physical, this is a good example of the manner in which a statistician can help to improve the accuracy of a laboratory technique.

17 LE ROY, H. L. **Statistische Methoden der Populationsgenetik.** Stuttgart: Birkhauser Verlag, 1960, pp. 397, 67.5 Swiss Francs.

C. HARTE, *Universität zu Köln, Germany.*

This book contains an extensive description of statistical methods used in modern population genetics. The reader is assumed to have some knowledge of genetics and general biometrical methods, especially correlation methods and the analysis of variance. The first chapter deals with the influence of environment, genotype, and the combination of both on the expression of the phenotype. First, statistical and biometrical models are constructed. Then statistical methods are explained, and the use of the formulae demonstrated by numerical examples. The cases discussed are simple allelism, multiple allelism and polygenic inheritance. On this basis the next chapter gives the statistical treatment of the interdependence of genetical and environmental factors. Here the method of path coefficients is introduced, and by this method the analysis of correlations and interactions for different degrees of relationship is explained.

The third chapter deals with the same problems, but here they are analysed by means of the analysis of variance for the cases of a hierarchical and a two-way classification (incomplete and complete analysis of variance). The author gives both the method of analysis and the interpretation of the analytical results. The methods apply mainly to animal breeding, but in addition the results are generalised by discussing the special problems which arise in plant breeding and human genetics.

In the fourth chapter the methods given in Chapters 2 and 3 are applied to the special case of artificial selection in respect of one or more characteristics. Here not only the classical designs but also modern methods are considered. The author compares regression analysis, the method of path coefficients and the analysis of variance. Different methods of treatment are also compared with respect to their efficiency and the amount of information they can give.

In all chapters the formulae which are used are derived in a valid manner and the methods of analysis are illustrated by good examples. In several places there are diagrams which help to make the complex connections quite clear. The methods of analysis are clearly illustrated by detailed numerical examples. At the end of the book there is an extensive list of literature, which not only contains the papers mentioned in the text, but is in fact a bibliography on all the questions raised in the course of the discussions. In this way it is easy for the reader to find references in his own field of work on special questions which are not given a place in the book. The last pages give, besides the index and a list of authors, three very instructive tables. Their two-way classification according to problems and methods makes it possible to see the connections between the chapters and to locate the questions which are discussed under different topics from different points of view.

The reader who knows enough in advance to follow the author and who has enough energy to work through nearlv 400 pages, will benefit greatly from this book.

18 LI, C. C. **Numbers from Experiments.**—A basic Analysis of Variation. Pittsburgh: The Boxwood Press 1959. pp. 1x + 106, Cloth $4.00, paper $2.75.

G. H. Jowett, *University of Melbourne, Australia.*

This book is mainly concerned with explaining the algebra involved in the analysis of variance applied to one-way, hierarchical and two-way classifications, and to linear regression.

The author is clearly fascinated by the partitioning of sums of squares and goes to considerable trouble to show how this happens. He makes considerable use of an interesting diagrammatic representation of a comparison between sums of squares as a comparison between areas made up of actual geometrical squares, which is apparently his own invention.

Summation and suffix notations are used freely, particularly in later chapters, and matrices and orthogonal contrasts are also used, but only to demonstrate the splitting up of a sum of squares in an alternative manner. No basic statistical theory is given, nor is any serious attempt made to explain other aspects of the mathematical statistics. Some space is devoted to establishing expectations of mean squares, but this is based on quoted theorems of the type,

$$E(y_1 + y_2 + \cdots) = E(y_1) + E(y_2) + \cdots ,$$

$$E(y_1y_2) = \mu^2 \qquad (y_1 , y_2 \text{ independent}),$$

the plausibility of which is justified mainly by reference to a very special case where y can take each value with equal probability. An attempt is also made to demonstrate the concept of a linear model; an ingenious example—"Grandma's birthday party"—is used for this purpose, deviations from observed means being interpreted as gains and losses in a poker game by the grandchildren, the money being provided by Grandma with some interference by Grandpa. Illustrative examples used in the exposition are deliberately artificial, being chosen to employ simple numbers and thus concentrate attention on relationships and procedure rather than to suggest applications.

Since the book does not cover the meaning and application of the techniques, it gives little help to anyone seeking to design or interpret the analysis of an experiment. On the other hand, it is not a suitable introduction to analysis of variance for a student of mathematical statistics or statistical method; it is loose and evasive about such important matters as degrees of freedom, testing hypotheses and independence, brings in sophisticated and inadequately explained jargon, and makes vague and unhelpful references to advanced mathematical topics (e.g. the rank of quadratic forms) in cases where the author has not been able to think of simple explanations. However, there is material here and there which, selected with discrimination, could give a teacher ideas for presentation or just tilt the balance of understanding for a student having difficulty with the algebra of the technique.

THE BIOMETRIC SOCIETY

MEETINGS OF E. N. A. R.

Annual Meeting, 1961

The annual meeting of the Eastern North American Region of the Biometrics Society is being held during the meetings of the American Statistical Association at the Roosevelt Hotel in New York City, December 27–30, 1961. Jointly with the Biometrics Section of A. S. A., ten sessions of invited papers and one session of contributed papers have been arranged.

Spring Meeting, 1962

E. N. A. R. will meet jointly with the Institute of Mathematical Statistics at the University of North Carolina, Chapel, Hill, April 12–14, 1962. M. E. Turner, Division of Biometry, Medical College of Viginia, and A. F. Bartholomay, Biophysics Laboratory, Harvard Medical School, Cambridge, are co-chairmen of the E. N. A. R. program committee. H. B. Wells, Department of Biostatistics, University of North Carolina, Chapel Hill is handling local arrangements for E. N. A. R.

Titles for contributed papers should be sent to M. E. Turner, who will then supply authors with abstract forms and appropriate instructions.

Annual Meeting, 1962

The annual meeting of E. N. A. R. in 1962 will be held during the meetings of the American Statistical Association at the Hotel Leamington in Minneapolis, September 7–10, 1962. Joint sessions with the Biometrics Section of A. S. A. are being arranged.

W. N. A. R.

MINUTES OF THE 1961 WNAR BUSINESS MEETING

The annual meeting of WNAR, Biometrics Society was held at 2:30 P. M. June 16, 1961 at the University of Washington, Seattle. William F. Taylor, President, WNAR, presided.

The Secretary's and Treasurer's reports were read and approved. The WNAR student award (consisting of a free year's subscription to *Biometrics*) was discussed at length and a motion was passed providing that any graduate student with a dual interest in the biological sciences and statistics should be considered eligible for the award.

A possible meeting with the AAAS at Denver December 26–31 was discussed and the general concensus of opinion was that we should participate in this meeting.

A motion was passed unanimously that WNAR meet for our annual meeting with AIBS at Oregon State University, Corvallis during the last week of August, 1962.

The nominating committee gave their report and nominations were made from the floor. Those nominated were: for President, Carl A. Bennett, for the two Regional Committee vacancies, Gordon E. Dickerson, Charles Mode, Henry Tucker and Alvin Wiggins.

The meeting adjourned at 3:20 P.M. Present at the meeting: Becker, Bennett, G., Bennett, C., Berry, Bohidar, Brown, Calvin, Chapman, Hopkins, Leitch, Mantel, Nash, Petersen, Puri, Russell, Sandomire, Taylor, Tucker, Vaughan, Wiggins.

WNAR ELECTION RESULTS

Carl A. Bennett was elected WNAR President for 1962–63 and Charles J. Mode and Henry Tucker were elected to the Regional Committee for 1962–64.

JOINT MEETING OF WNAR AND AAAS

The Western North American Region will meet with the American Association for the Advancement of Science in Denver on December 28, 1961. There will be three sessions entitled "Experimental Design," "Estimation of Populations (mobile and others)," and a "Contributed Papers Session." Members wishing to present papers should contact Dr. Franklin A. Graybill, Statistical Laboratory, Colorado State University, Fort Collins, Colorado, Program Chairman.

CHANGES OF MEMBERSHIP
(July 15-October 15, 1961)

Changes of Address

Dr. Daniel J. Baer, 519 Chaucer Road, Dayton, Ohio, U.S.A.

Mr. C. G. Barraclough, Chemistry Department, University of Melbourne, Parkeville N. 2, Victoria, Australia

Dr. Austin W. Berkeley, Department of Psychology, Boston University, 700 Commonwealth, Boston 15, Massachusetts, U.S.A.

Mr. Paul V. Blair, Institute of Enzyme Research, 1702 University Avenue, University of Wisconsin, Madison 5, Wisconsin, U.S.A.

Mr. Andre Bodeaux, Faculte Agronomique, Usumbura, Ruanda-Urundi, Belgian Congo

Mr. Neeti R. Bohidar, Statistical Laboratory, Utah State University, Logan, Utah, U.S.A.

Mr. Philippe F. Bourdeau, 370 Prospect Street, New Haven 11, Connecticut, U.S.A.

Mr. N. Charliers, 199 Bonneville Sclayn, Province de Namur. Belgium

Dr. J. B. Chassan, Hoffman-LaRoche, Nutley 10, New Jersey, U.S.A.

Mr. Pan Leang Cheav, 1 rue Lucien Petit, Gembloux, Belgium

Mr. William P. Chu, 5824 Fairfax Avenue, South Minneapolis, Minnesota. U.S.A.

Dr. Ellsworth B. Cook, American Society Pharmacology and Experimental Therapeutics, 9650 Wisconsin Avenue, NW, Washington 14, D. C., U.S.A.

Mr. Tiberius Cunia, 2835 McCarthy Street, City of St. Laurent, Quebec, Canada

Mr. R. N. Curnow, Unit of Biometry, The University, Reading, Berkshire, England

Mr. Pierre T. A. Dagnelie, 53 Chaussee de Charleroi, Gembloux, Belgium

Mr. Marc Dalebroux, 74 Planoy, Biesme-lez, Belgium

Mr. R. de Coene, 79 rue Lincoln, Bruxelles 18, Belgium

Prof. Martin E. Dehousse, 88 avenue Maeterlinck, Brussels 3, Belgium

Dr. B. Diamantis, 2129 Scudder Street, St. Paul 8, Minnesota, U.S.A.

Mr. H. M. Dicks, South Africa Sugar Association, Experiment Station, P. O. Mt. Edgecombe, Natal, South Africa

Miss Martha W. Dicks, Division of Home Economics, Box 3354. University Station, Laramie, Wyoming, U.S.A.

Mr. Victor A. Dirks, 12800 Dupont Avenue, S, Savage, Minnesota, U.S.A.

Dr. N. R. Draper, Department of Statistics, University of Wisconsin, Madison, Wisconsin, U.S.A.

Mr. Bruce A. Drew, The Pillsbury Company, 311 Second Street, S.E., Minneapolis 14, Minnesota, U.S.A.

Mr. L. Lee Eberhardt, Box 562, Concord, California, U.S.A.

Mr. Harvey Eisenberg, 3722 Cedar Drive, Lochearn, Baltimore 7, Maryland, U.S.A.

Dr. Khalil M. El-Kashlan, 16 Green Street, Moharrem Bey, Alexandria, Egypt

Mr. Roger Firmin, 10 avenue R. Neybergh, Bruxelles, Belgium

Mr. Andris Fogelmanis, 914 Sunset Drive, Pacific Grove, California, U.S.A.

Mr. J. Fraselle, rue des Bas-Pres, Salzinnes (Namur), Belgium

Dr. Seymour Geisser, Section on Biometry, Nat. Inst. of Arthritis and Metabolic Diseases, Bethesda, Maryland, U.S.A.

Dr. Luc Goeminne, 6 Bagattenstraat, Gand, Belgium

Mr. Richard A. Greenberg, 430 Central Avenue, New Haven, Connecticut, U.S.A.

Mr. Samuel W. Greenhouse, Section on Statistics and Mathematics, National Institute of Mental Health, Bethesda, Maryland, U.S.A.

Mrs. Fredrica G. Halligan, Warwick Towers, 9 Lafayette Court, Greenwich, Connecticut, U.S.A.

Dr. Emil L. Gumbel, Columbia University, 413 West 117 Street, New York 27, N. Y., U.S.A.

Prof. John Gurland, U.S. Army Math. Research Center, University of Wisconsin, Madison, Wisconsin, U.S.A.

Mr. M. A. Guzman, 10 Dixie Trail, Raleigh, North Carolina, U.S.A.

Mr. A. M. Haider, Federation of Industries, Ottoman Bank Building, Baghdad, Iraq

Dr. W. D. Hanson Department of Genetics, North Carolina State College, Raleigh, North Carolina, U.S.A.

Mr. Jacques Hardouin, 35 Via San Girolama, Perugia, Italy

Mr. J. A. G. Hemptinne, 29 rue A. de Wasseige, Wepion, Fooz, Belgium

Professor Dr. Leon Hennaux, Ville Beau Sejour, Coril Noirmont (Brabant) Belgium

Mr. Jean Henry, 3 avenue Copyn Malaise, La Hulpe, Belgium

Mr. John R. Howell, Route 5, Box 364, Orlando, Florida, U.S.A.

Dr. Pierre O. Hubinont, 25 avenue Ad. Demeur, Bruxelles 6, Belgium

Mr. Paul V. Hurt, 4721 Ames Street, Madison 5, Wisconsin, U.S.A.

Mr. Floribert A. L. G. Jurion, 8 avenue Isidore Gerard, Bruxelles 16, Belgium

Dr. Richard A. Lang, Department of Statistics, P. O. Box 5457, Raleigh, North Carolina, U.S.A.

Dr. Lonnie L. Lasman, 124 Claridge, Eau Gallie 7, Florida, U.S.A.

Dr. Jean Lebrun, 12 av. des Lucioles, Brussels, Belgium

Mr. G. H. Lion, 61 rue Gabrielle, Bruxelles 18, Belgium

Mr. Donald G. MacEachern, 1717 Brook Avenue, S.E., Minneapolis 14, Minnesota, U.S.A.

Mr. John W. Mayne, 654 Sherbourne Road, Ottawa 3, Ontario, Canada.

Mr. James H. Meade, Jr., Institute of Statistics, P. O. Box 5457, Raleigh, North Carolina, U.S.A.

Mr. Hyman Menduke, 7808 Haines Road, Cheltenham, Pennsylvania, U.S.A.

Mr. Forest L. Miller, 111 Villanova Road, Oak Ridge, Tennessee, U.S.A.

Dr. Sigeiti Moriguti, Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo, Japan

Mr. Floyd R. Olive, McKamie, Arkansas. U.S.A.

Mr. James G. Osborne. Old Chapel Hill Road, Route 1, Box 97, Durham, North Carolina, U.S.A.

Mr. Jyun Otsuka, Kaname-machi 14, Toshimaku, Tokyo, Japan

Miss Florence E. Petzel, Department of Home Economics, University of Texas, Austin, Texas, U.S.A.

Mr. Andre Pieters, 39 Eeklastraat, Mariakerke (Gent) Belgium

Mr. Jean Reuse, Clos des 4 Vents. Rue de Moorsel, (Brabant) Tervueren, Belgium

Mr. Julien F. M. Ronchaine, 70 Grand Rue, Gembloux, Belgium

Mr. Charles E. Rossiter, Pneumoconiosis Research Unit, Llandough Hospital, Penarth, Glamorgan, Wales

Dr. Francesco Sella, Radiation Committee, Room 3468, United Nations. New York, N. Y., U.S.A.

Mr. A. B. Siegelaub, 721 Walton Avenue, New York 21, N. Y., U.S.A.

Mr. William A. Small, P. O. Box 122-A, Tennessee Polytechnic Institute, Cookeville. Tennessee, U.S.A.

Dr. H. Fairfield Smith, FAO-UN, P. O. Box 1555, Teheran. Iran

Dr. John H. Smith, Department of Mathematics and Statistics, American University, Washington 16, D.C., U.S.A.

Mrs. Grace Scholz Spitz, 803 Marlo Drive, Falls Church, Virginia, U.S.A.

Mr. Francois Sterckx, Aartcentrum 66, Geel, Belgium

Dr. Howard L. Stier, United Fruit Company, 30 St. James Avenue, Boston 16, Massachusetts, U.S.A.

Dr. Robert J. Taylor, CCNCS-NCI, National Institutes of Health, Bethesda, Maryland, U.S.A.

Prof. Robert S. Temple, College of Agriculture, University of Tennessee, Knoxville, Tennessee, U.S.A.

Mr. George W. Thomson, 15093 Faust Avenue, Detroit 23, Michigan, U.S.A.

Mr. M. Ulehla, 8 Carlysle Street, E. Hawthorn. Victoria, Australia

Prof. A. van den Hende. 233 Coupure Links, Gand, Belgium

Miss Cecile Van Kerchove, 13 rue Archimede, Brussels, Belgium

Mrs. Pearl A. Van Natta, 1303 South Eudora, Denver 22, Colorado, U.S.A.

Dr. B. Veen, van Tienhovenlaan 13, Velp, Netherlands

Mr. Glen F. Vogel, 6600 Luzon Avenue, NW. Washington 12, D. C., U.S.A.

Prof. Maurice Welsch, 99 Rue General Jacques, Embourg, Belgium

Dr. R. Wette, M. D. Anderson Hospital, 6723 Bertner Avenue, Houston 25, Texas, U.S.A.

Mr. Robert F. White, Science Information Lab., Smith, Kline and French, Philadelphia, Pennsylvania, U.S.A.

Dr. Manfred Woelke, Killicklaan 689, Les Marais, Pretoria, South Africa

New Members

At Large

Mr. A. M. Haider, Federation of Industries, Ottoman Bank Building, Baghdad, Iraq

Dr. Helmut V. Muhsam, Professor of Statistics, Hebrew University, Jerusalem, Israel

Australia

Mr. P. J. Brockwell, Department of Statistics, University of Melbourne, Victoria, Australia

Mr. Peter D. Finch, Department of Statistics, University of Melbourne, Victoria, Australia

Miss Betty Laby, Department of Statistics, University of Melbourne, Victoria, Australia

Mr. J. S. Maritz, Department of Statistics, University of Melbourne, Victoria, Australia

Miss Ilze Raudzins, 36 Boulder Avenue, Redcliffe, W. Australia

Mr. K. V. Richardson, Department of Statistics, University of Melbourne, Victoria, Australia

ENAR

Mr. Charles Anello, 5905 Montgomery Street, Baltimore 7, Maryland, U.S.A.

Mr. Charles E. Antle, 1620 University Avenue, Stillwater, Oklahoma, U.S.A.

Mr. Thomas P. Bogyo, Department of Experimental Statistics, North Carolina State College, Raleigh, North Carolina, U.S.A.

Mr. DuWayne C. Englert, Population Genetics Institute, Purdue University, Lafayette, Indiana, U.S.A.

Prof. George A. Ferguson, Department of Psychology, McGill University, Montreal, Canada

Mr. R. E. Foster, Forest Pathology Investigations, Department of Forestry, Victoria, B. C., Canada

Mrs. Iris M. Kiem, Department of Medicine, University of Miami, Miami 36, Florida, U.S.A.

Miss Joan Klebba, 1469 South 28th Street, Arlington 6, Virginia, U.S.A.

Mr. James L. Pate, Department of Psychology, University of Alabama, University, Alabama, U.S.A.

Dr. Norman B. Rushforth, Developmental Biology Center, Western Reserve University, Cleveland, Ohio, U.S.A.

Mr. Jack Sebring, 1515 Ogden Street, NW, Washington 10, D. C., U.S.A.

Germany

Dr. O. Dittmar, Inst. f. Forstwiss, Alfred-Moeller-Str., Eberswalde, Germany

Sweden

Prof. Dr. E. Effenberger, (Hamburg 13) Gartenstrasse 2, Alsterchaussee, Erfurt, Sweden

Switzerland

Dr. Klaus G. Konig, Zahnarztliches Institut der Universitat, Zurich 28, Postfach, Switzerland

Dr. Thomas M. Marthaler, Dental Institute, University of Zurich, Postfach, Zurich 28, Switzerland

Dr. Heinrich Wagner, Ungargasse 32, Wien III, Austria

Dr. K. Wuhrmann, Hofstr. 139, Zurich 44, Switzerland

WNAR

Dr. G. Eric Bradford, Department of Animal Husbandry, University of California, Davis, California, U.S.A.

Mr. Howard F. Gingerich, 1817½ Berkeley Way, Berkeley 3, California, U.S.A.

Mr. Melville R. Klauber, Bldg. 315, Stanford Village, Stanford, California, U.S.A.

Dr. Orsell M. Meredith, Nuclear Medicine and Radiation Biology, UCLA Medical School, Los Angeles 24, California, U.S.A.

Mr. William L. LeStourgeon, Statistics Group, Los Alamos Scientific Labs., Los Alamos, New Mexico, U.S.A.

Dr. Daniel G. Siegel, Atomic Bomb Casualty Commission, Navy 3912, FPO, San Francisco, California, U.S.A.

Dr. Ervin P. Smith, Animal Science Department, Montana State College, Bozeman, Montana, U.S.A.

Mr. N. Scott Urquhart, Computing Center CSURF, Colorado State University Fort Collins, Colorado, U.S.A.

NEWS AND ANNOUNCEMENTS

Members are invited to transmit to their National or Regional Secretary (if members at large, to the General Secretary) news of appointments, distinctions, or retirements, and announcements of professional interest.

NEWS ABOUT MEMBERS

Jacob B. Chassan has been appointed statistician for the Research Department of Hoffmann-La Roche, Inc., at Nutley, N. J.

C. Philip Cox, formerly of the N.I.R.D., Shinfield, England, has joined the staff of the Statistical Laboratory and Department of Statistics, Iowa State University, as Associate Professor. Professor Cox will be teaching and conducting research in the area of biological statistics.

Seymour Geisser has been recently appointed Chief of Biometry of the National Institute of Arthritis and Metabolic Diseases.

Dewey L. Harris, has been appointed as Assistant Professor to the staff of the Statistical Laboratory and Department of Statistics, Iowa State University. Dr. Harris will be consulting, teaching and doing research in the area of Genetics Statistics.

Donald H. Hazelwood, formerly NIH Fellow, Department of Zoology, Washington State University has taken a position as Assistant Professor of Zoology, University of Missouri.

Donald K. Hotchkiss, has been appointed as Assistant Professor to the staff of the Statistical Laboratory and Department of Statistics, Iowa State University. Dr. Hotchkiss will be consulting, teaching, and doing research in the area of statistics applied to animal nutrition.

Robert Macey is now Assistant Professor, Department of Physiology, University of California, Berkeley, moving from the University of Illinois School of Medicine.

J. N. K. Rao of the Statistical Laboratory, and Department of Statistics, Iowa State University, has been promoted to rank of Assistant Professor. Dr. Rao will be teaching and conducting research in survey sampling in his new assignment.

Elmer E. Remmenga, Chief, Computing Center, Colorado State University, will be on sabbatical leave from July, 1961 to July, 1962.

Patrick K. Tomlinson has recently become an Aquatic Biologist, Ocean Salmon Investigation, California Department of Fish and Game. He formerly was in the Biostatistics Section.

Richard W. Vail, Jr. has taken a position as statistician in the Aerospace Corporation, Florida. He previously was statistician for Aerojet General Corporation, Azusa, California.

Pearl A. Van Natta, formerly Biometrician at the Child Research Council, Denver, Colorado, is a self employed statistical consultant.

SOUTHERN METHODIST UNIVERSITY—NEW PROGRAM IN STATISTICS

Southern Methodist University announces the opening of a new Department of Mathematical and Experimental Statistics and Statistical Laboratory in the

Graduate School. Faculty members include Drs. Paul D. Minton and Vanamamalai Seshadri, assisted part-time by Mr. Del West and Mr. Mack Usher.

Coursework leading to the degree of Master of Science will be offered, including courses in mathematical statistics, experimental statistics, probability, sampling theory and design of experiments.

Departmental assistantships are available.

For further information, write to Dr. Paul D. Minton, Department of Mathematical and Experimental Statistics, Southern Methodist University, Dallas 22, Texas.

FLORIDA STATE UNIVERSITY—EXPANDED PROGRAM IN STATISTICS

Beginning in September 1962, the Department of Statistics at the Florida State University will expand its graduate program to study and research leading to the Doctor of Philosophy degree in statistics. The success of the present program, currently limited to study to the Master of Science degree in statistics, has led to the approval of the new advanced training. The curriculum will be modified and expanded to include advanced work in statistical inference and decision theory, stochastic processes, advanced probability, multivariate analysis, operations research, special topics in biometry, theory of general linear hypotheses, non-parametric statistics and sequential analysis. Advanced courses will be offered in alternate years. The program leading to the Master of Science degree will be modified to permit both the thesis and non-thesis types of programs.

The new program will be assisted through the recent addition to the staff of Dr. Vincent Hodgson from the London School of Economics and Political Science. Dr. Hodgson joints Drs. Ralph A. Bradley, Frank Wilcoxon, Richard G. Cornell and S. K. Katti in the Department of Statistics. Additional faculty appointments within the next two years are anticipated.

Facilities have been greatly improved through completion of a new mathematical sciences building in which instructional, laboratory and office space was designed for use of the faculty and graduate students in statistics. The new building also houses an IBM 709 computer for research use.

The Department has a limited number of teaching and research assistantships available. Proposals for three-year graduate fellowships are pending and such fellowships should be available for graduate students entering the new program in September. Interested students are invited to write to Dr. R. A. Bradley, Department of Statistics, the Florida State University, Tallahassee, Florida for further information.

CORRECTIONS

Chin Long Chiang [1960]. A Stochastic Study of the Life Table and Its Applications: I. Probability Distributions of the Biometric Functions. *Biometrics 16*, 618–35.

In formula (2), page 621, for 1_x read l_x. In formula (9), page 622, n should be deleted.

Chin Long Chiang [1961]. A Stochastic Study of the Life Table and Its Applications: III. The Follow-up Study with the Consideration of Competing Risks. *Biometrics 17*, 57–78.

In line 1, page 62, m should be replaced by n. In formula (21), page 66, for $S^2_{e\,\alpha}$ read $S^2_{\hat{e}\,\alpha}$. In formula (32), page 69, for $Q_{xk\ 12}$ read $Q_{xk.12}$. In formulas (57) and (58), page 73, the super 2 should be deleted from σ^2. In formula (58), page 73, for Λ_{xk} read Λ_{hk} . In column (2) of table 4, page 77, for 835.87 read 935.87. In line 5 from the bottom, page 77, for $\hat{Q}_{xk}$ read $\hat{q}_{xk}$.

J. K. Abraham [1960]. 154 Note: On an Alternative Method of Computing Tukey's Statistic for the Latin Square Model. *Biometrics 16*, 686–691.

In Table 3, cn^4 should multiply each $\sum^i$, $\sum^j$ and $\sum^k$ appearing, and in Table 4, beneath "Treatments", $\sum^k$ should be similarly corrected.

BIOMETRIE—PRAXIMETRIE

Tome I, N° 3–4 *Juillet-Octobre, 1960*

Tome II, N° 1 *Mars, 1961*

Publishers for the

SOCIETE ADOLPHE QUETELET
7 rue Héger-Bordet, Brussels 1, Belgium

Presses Universitaires de Bruxelles
50, av. Franklin Roosevelt
Brussels